



UNIVERSITY OF THESSALY

SCHOOL OF SCIENCE

MSC INFORMATICS AND
COMPUTATIONAL BIOMEDICINE

Machine Learning for Predicting Antimicrobial Resistance

MASTER DISSERTATION

Author:

Daniilidis Ap. Konstantinos  (R.N.: M2022049)

Supervisor:

Hatzigeorgiou G. Artemis, *Professor*

Lamia, 2024



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Δ.Π.Μ.Σ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΪΑΤΡΙΚΗ

Μηχανική Μάθηση για την Πρόγνωση της Αντιμικροβιακής Ανθεκτικότητας

Μεταπτυχιακή Εργασία

Συγγραφέας:

Δανιηλίδης Απ. Κωνσταντίνος  (Α.Μ.: M2022049)

Επιβλέπουσα:

Χατζηγεωργίου Γ. Άρτεμις, Καθηγήτρια

Λαμία, 2024

*"The successful man is the one
who finds out what is the matter with
his business before his competitors do."*

— *Alexander Fleming*

*"Science knows no country, because
knowledge belongs to humanity, and is
the torch which illuminates the world."*

— *Louis Pasteur*

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις¹, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.
2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.
3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια.
4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: **01/07/2024**

Ο Δηλών,

.....

Δανιηλίδης Απ. Κωνσταντίνος
Δ.Π.Μ.Σ Πληροφορική και Υπολογιστική Βιοϊατρική,
Πανεπιστήμιο Θεσσαλίας

Copyright © University of Thessaly [2024]
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

¹ «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Machine Learning for Predicting Antimicrobial Resistance

Author:

Daniilidis Ap. Konstantinos 

Three-member committee:

Hatzigeorgiou G. Artemis,

(Supervisor) Professor of the Department of Computer Science and Biomedical Informatics of the School Of Sciences of the University of Thessaly

Bagos G. Pantelis,

Professor of the Department of Computer Science and Biomedical Informatics of the School Of Sciences of the University of Thessaly

Braliou G. Georgia,

Assistant Professor of the Department of Computer Science and Biomedical Informatics of the School Of Sciences of the University of Thessaly

CONTENTS

	Page
List of Schemes	v
List of Tables	vi
List of Abbreviations	vii
Acknowledgements	viii
Abstract	ix
Περίληψη	x
1. Introduction	1
1.1. A Little Antibiotic History	1
1.2. Statistical Overview of Antimicrobial Resistance	3
1.2.1. Antimicrobial Resistance in Europe and European Economic Area	3
1.2.2. Antimicrobial Resistance Landscape in Greece	3
1.2.3. Antimicrobial Resistance Landscape in the United States	4
1.2.4. Antimicrobial Resistance and Covid-19 Relation	4
1.3. Modes of Antibiotics	4
1.4. Genetics Of Resistance	5
1.5. Ecological Roles of Antimicrobial Resistance	7
1.6. Nosocomial Infections	7
1.7. Antimicrobial Susceptibility Testing (AST)	8
1.8. Antimicrobial Resistance Databases	9
1.8.1. National Database of Antibiotic Resistant Organisms	9
1.8.2. Bacterial and Viral Bioinformatics Resource Center	10
1.8.3. National Antimicrobial Resistance Monitoring System	10
1.9. Machine Learning Algorithms	11
2. Methods	13
2.1. Public Data Collection	13
2.2. Snakemake Pipeline for Genomic Data Analysis	14
2.3. Assemblies Quality Control	15
2.4. Antibigram Tables	15
2.5. Antibigrams Quality Control	16
2.6. Annotation Pipeline for Analysis of WGS Data	17
2.6.1. Comprehensive Antibiotic Resistance Database (CARD)	17

2.6.2.	Antimicrobial Resistance Gene Finder Plus (AMRFinderPlus)	17
2.6.3.	Antimicrobial peptides (AMPs) Clustering	18
2.6.4.	AMPs Database and Alignment	19
2.6.5.	Determination of Non-Translated Genes	19
2.7.	Building and Evaluating Machine Learning Models	20
2.7.1.	Encoding Methods	21
2.7.1.1.	k -mer Encoding	21
2.7.1.2.	One-hot Encoding	22
2.7.1.3.	PSSM Encoding	22
2.7.2.	XGBoost Model Training and Testing	23
2.7.3.	Random Forest Training and Testing	23
2.7.4.	Hyperparameters Tuning	24
2.7.5.	Performance Metrics in Antimicrobial Resistance Prediction	24
2.7.6.	k -fold Cross – Validation	26
3.	Results	27
3.1.	Data Collection and Pre-processing	27
3.2.	k -fold Cross – Validation	27
3.3.	XGBoost Performance Metrics	29
3.4.	Random Forest Performance Metrics	33
4.	Discussion	37
5.	Conclusion	40
6.	Availability of Data	41
	Bibliography	42

LIST OF FIGURES

1. Introduction	
1.1 Timeline of Key Antibiotic Resistance Events	2
1.2 The image illustrates the various mechanisms of action of antibacterial drugs	6
1.3 Antimicrobial Susceptibility Testing (AST) Process	9
2. Methods	
2.1 Quality Control Analysis of Genomic Assemblies by Bacterial Genus	16
2.2 Distribution of Gentamicin Phenotypes	21
3. Results	
3.1 Gentamicin predicted Phenotypes for k -mer size 3, using XGBoost model and Cross-Validation	28
3.2 Gentamicin predicted Phenotypes for k -mer size 4, using XGBoost model and Cross-Validation	28
3.3 Gentamicin predicted Phenotypes for k -mer size 5, using XGBoost model and Cross-Validation	29
3.4 Gentamicin predicted Phenotypes for k -mer size 3, using the XGBoost model	30
3.5 Gentamicin predicted Phenotypes for k -mer size 4, using the XGBoost model	30
3.6 Gentamicin predicted Phenotypes for k -mer size 5, using the XGBoost model	31
3.7 Performance Metrics for XGBoost Model in Predicting Antimicrobial Resistance using <code>softprob</code> objective function	32
3.8 Gentamicin predicted Phenotypes using XGBoost algorithm and one-hot encoding	33
3.9 Gentamicin predicted Phenotypes using XGBoost algorithm and PSSM encoding	33
3.10 Gentamicin predicted Phenotypes for k -mer size 3, using Random Forest algorithm	34
3.11 Gentamicin predicted Phenotypes for k -mer size 4, using Random Forest algorithm	34
3.12 Gentamicin predicted Phenotypes for k -mer size 5, using Random Forest algorithm	35
3.13 Gentamicin predicted Phenotypes using Random Forest algorithm and one-hot encoding	35
3.14 Gentamicin predicted Phenotypes using Random Forest algorithm and PSSM encoding	36

LIST OF TABLES

2.1	Description of Columns in the Antibigram Data Frame	16
2.2	CD-HIT results using different clustering parameters	18
2.3	Bacterial strains used to derive 5S, 16S and 23S rRNA copies	20
3.1	k -mer data set used with stratified k -fold cross-validation on the XGBoost model	29

LIST OF ABBREVIATIONS

MDR	Multi-drug Resistance
AMR	Antimicrobial Resistance
HGT	Horizontal Gene Transfer
EUCAST	European Committee on Antimicrobial Susceptibility Testing
WGS	Whole Genome Sequencing
DALYs	Disability-adjusted life-years
EEA	European Economic Area
ECDC	European Centre for Disease Prevention and Control
CDC	Center for Disease Control and Prevention
NARMS	National Antimicrobial Resistance Monitoring System
NDARO	National Database of Antibiotic Resistant Organisms
NCBI	National Center for Biotechnology Information
BV-BRC	Bacterial and Viral Bioinformatics Resource Center
ARGs	Antimicrobial Resistant Genes
HAI s	Hospital-acquired Infections
CARD	Comprehensive Antibiotic Resistance Database
ML	Machine learning
MIC	Minimum Inhibitory Concentration
SVM	Support Vector Machine
ICU	Intensive Care Unit
CNN s	Convolutional Neural Networks
AST	Antimicrobial Susceptibility Testing
MIC	Minimum Inhibitory Concentration
CLSI	Clinical and Laboratory Standards Institute
CSV	Comma-separated Values
ARO	Antibiotic Resistance Ontology
AMP s	Antimicrobial peptides
BLAST	Basic Local Alignment Search Tool
PSSM	Position - Specific Scoring Matrix
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
FPR	False Positive Rate
ANN s	Artificial Neural Networks

ACKNOWLEDGEMENTS

This thesis was conducted from January 2024 to July 2024 to fulfill the requirements for the Master's degree from the Department of Computer Science and Biomedical Informatics at the University of Thessaly. Completing this research marks a significant milestone in my academic journey, representing the culmination of my postgraduate studies and the transition to a more advanced research orientation in the field of biomedical informatics. Moreover, the experience gained from this research has been immensely enriching, providing me with a deeper understanding of data set creation and machine learning applications and enhancing my skills in data analysis, algorithm implementation, and collaborative research. This journey has not only broadened my technical expertise, but also fostered critical thinking and problem-solving abilities, essential for advancing in the field of bioinformatics.

This thesis was implemented with the support of several individuals, and it is important to express my heartfelt gratitude to them. I would like to express special thanks to my supervisor, Professor Artemis Hatzigeorgiou, for her exceptional guidance, support, and the trust she placed in me. Her mentorship was invaluable, providing both direction and encouragement throughout the research process, and our collaboration has left me with many positive and lasting impressions. Additionally, I am deeply grateful to Argyris Skoulakis, a PhD candidate at the DNA Intelligent Analysis Laboratory (Diana Lab), for his significant contributions and unwavering support. His expertise and our productive collaboration from the outset of this project were instrumental in its successful completion.

ABSTRACT

The problem of Antimicrobial Resistance (AMR) is one of the most important public health problems in the world, which poses a great threat to modern medicine. The development of microbial resistance to antibiotics is a natural phenomenon that occurs when bacteria encounter antibiotics. This particular resistance can be transmitted between bacteria, through the exchange of plasmids or genetic material, and furthermore, it can lead to serious infections that cannot be treated with ordinary antibiotics. However, conventional antimicrobial susceptibility testing is time-consuming, low-throughput, and requires microbial cultures and testing with different antibiotics to identify the microbial phenotype.

On the other hand, in recent years, access to Next Generation Sequencing experiments, as well as the development of innovative bioinformatics tools and databases, has led to better clinical decision-making through the identification of microbial resistance genes. However, improving these diagnostic tools is a problem that is open to further development as the available genetic data increases and the accuracy of machine learning algorithms continues to improve.

The aim of this master dissertation is initially the construction of a dataset of bacterial genomes, which, after bioinformatic analysis, can be annotated in a specific way, in order to identify possible mutations, genes related to microbial resistance, as well as other sequences that contribute to phenotype. Finally, these data are fed into various machine learning algorithms, and antimicrobial resistance is predicted, achieving interesting conclusions for each bacterial species.

Keywords— Antimicrobial Resistance, Bacteria, Antibiotics, Machine Learning, Genomic Data, Clinical Diagnostics

ΠΕΡΙΛΗΨΗ

Το πρόβλημα της αντιμικροβιακής ανθεκτικότητας στα αντιβιοτικά είναι ένα από τα σημαντικότερα προβλήματα δημόσιας υγείας στον κόσμο, το οποίο αποτελεί μεγάλη απειλή για τη σύγχρονη ιατρική. Η ανάπτυξη μικροβιακής ανθεκτικότητας, είναι ένα φυσικό φαινόμενο που συμβαίνει όταν τα βακτήρια αντιμετωπίζουν αντιβιοτικά. Η εν λόγω ανθεκτικότητα, μπορεί να μεταδοθεί μεταξύ των βακτηρίων, μέσω της ανταλλαγής πλασμιδίων ή γενετικού υλικού και επιπλέον, μπορεί να οδηγήσει σε σοβαρές λοιμώξεις που δεν μπορούν να θεραπευτούν με τα συνηθισμένα αντιβιοτικά. Ωστόσο, οι συμβατικές δοκιμές αντιμικροβιακής ευαισθησίας είναι χρονοβόρες, μικρής διεκπεραιωτικής ικανότητας και απαιτούν μικροβιακές καλλιέργειες και έλεγχο με διαφορετικά αντιβιοτικά, ώστε για γίνει εντοπισμός του μικροβιακού φαινοτύπου.

Από την άλλη πλευρά, τα τελευταία χρόνια, η προσπάθεια σε πειράματα Αλληλούχησης Επόμενης Γενιάς, καθώς και η ανάπτυξη καινοτόμων βιοπληροφορικών εργαλείων και βάσεων δεδομένων, οδήγησε στην καλύτερη λήψη κλινικών αποφάσεων, μέσω της ταυτοποίησης των γονιδίων μικροβιακής ανθεκτικότητας. Ωστόσο, η βελτίωση αυτών των διαγνωστικών εργαλείων, είναι ένα πρόβλημα που επιδέχεται περαιτέρω ανάπτυξης, καθώς τα διαθέσιμα γενετικά δεδομένα αυξάνονται και η ακρίβεια των αλγορίθμων μηχανικής μάθησης ολοένα και βελτιώνεται.

Στόχος της παρούσας μεταπτυχιακής εργασίας, είναι αρχικά η κατασκευή ενός συνόλου δεδομένων (dataset) βακτηριακών γονιδιωμάτων, το οποίο κατόπιν βιοπληροφορικής ανάλυσης, δύνανται να σχολιαστεί με συγκεκριμένο τρόπο, ώστε να εντοπιστούν πιθανές μεταλλάξεις, γονίδια που σχετίζονται με τη μικροβιακή ανθεκτικότητα, καθώς και άλλες αλληλουχίες που συμβάλλουν στο φαινότυπο. Τέλος, τα δεδομένα αυτά τροφοδοτούνται σε διάφορους αλγορίθμους μηχανικής μάθησης και γίνεται πρόγνωση της αντιμικροβιακής ανθεκτικότητας, καταλήγοντας σε ενδιαφέροντα συμπεράσματα για κάθε βακτηριακό είδος.

Λέξεις Κλειδιά: Αντιμικροβιακή Ανθεκτικότητα, Βακτήρια, Αντιβιοτικά, Μηχανική Μάθηση, Γονιδιακά Δεδομένα, Κλινική Διαγνωστική

1 | INTRODUCTION

It was 1928, when Alexander Fleming discovered penicillin, revolutionizing medicine forever (Fleming, 1944). His chance observation of mold inhibiting bacterial growth in a petri dish, paved the way for one of the most significant medical breakthroughs of the 20th century. Fleming's discovery laid the foundation for the development of antibiotics, transforming the treatment of bacterial infections and saving countless lives. The impact of Penicillin resonates to this day, underscoring the importance of scientific curiosity and perseverance in the face of uncertainty.

Since then, dozens of antibiotics have been discovered, continuing Fleming's legacy and expanding the arsenal against infectious diseases. These antibiotics, derived from various sources including fungi, bacteria, and synthetic compounds, have become indispensable tools in modern medicine for preventing or treating infections (Ventola, 2015). From penicillins to cephalosporins, tetracyclines to macrolides, these medications target a wide range of bacteria, combating infections ranging from minor skin ailments to life-threatening illnesses (Julian Davies and D. Davies, 2010). However, their widespread use has also led to the emergence of antibiotic resistant strains of bacteria, highlighting the ongoing need for judicious antibiotic stewardship and the development of novel therapies to combat evolving microbial threats (World Health Organization, 2022; Ventola, 2015).

1.1. A LITTLE ANTIBIOTIC HISTORY

The term "antibiotic" was first used in 1942 by Selman Waksman and his collaborators to describe substances produced by microorganisms that are antagonistic to the growth of other microbes in high dilution (Aminov, 2010). However, the history of antibiotics dates back to ancient times when various civilizations, such as the Egyptians and Greeks, applied moldy bread to infected wounds, unknowingly using the antibacterial properties of molds. It was not until the late 19th century that scientists began to observe antibacterial chemicals in action. Paul Ehrlich, a German physician, discovered that certain chemical dyes colored some bacterial cells but not others, leading to the development of a chemical called arsphenamine in 1909, an effective treatment for syphilis and the first modern antibiotic (Aminov, 2010).

The discovery of penicillin by Alexander Fleming in 1928 marked the beginning of the golden age of antibiotics. Fleming noticed that a fungus, *Penicillium notatum*, had contaminated a culture plate of Staphylococcus bacteria and created bacteria-free zones wherever it grew (Fleming, 1944). This accidental discovery led to the isolation and mass production of penicillin, which proved to be a 'wonder drug' during World War II, saving countless lives. The success of penicillin spurred the search for other naturally occurring antibiotics, leading to the discovery of over 20 antibiotics by Waksman himself (Gould, 2016).

Following Fleming's groundbreaking discovery, the 1940s and 1950s became known as the golden era of antibiotic discovery. During this time, many new antibiotics were identified and developed from soil bacteria and fungi (Figure 1.1), significantly expanding the arsenal available to fight bacterial infections. Streptomycin, discovered in 1943 by Albert Schatz, a graduate student working in Selman Waksman's laboratory, became the first effective treatment for tuberculosis and marked a major milestone in medical history (Schatz *et al.*, 2005). Additionally, other antibiotics such as chloramphenicol, tetracyclines, and erythromycin were discovered, each with unique properties that broadened the spectrum of treatable infections and reduced the mortality rate of numerous diseases.

The development and widespread use of antibiotics revolutionized medicine and led to dramatic improvements in public health. However, it also set the stage for the emergence of antibiotic resistance, a significant challenge that continues to impact healthcare today. As bacteria evolve and develop resistance to existing antibiotics, the need for novel antibiotics and alternative treatment strategies remains critical, emphasizing the ongoing importance of research and innovation in the field of antimicrobial therapy (Aminov, 2010).

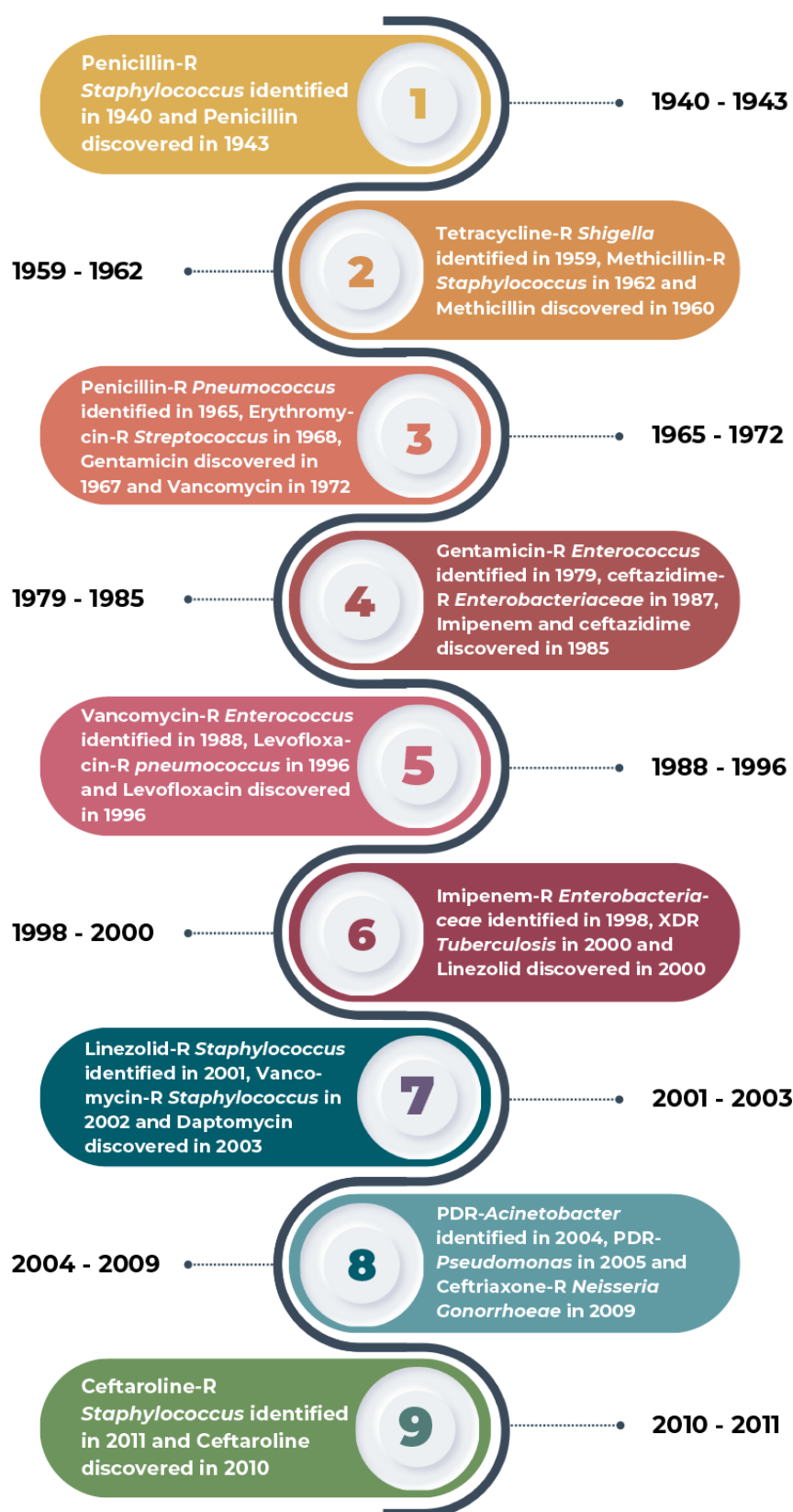


Figure 1.1 | Timeline of Key Antibiotic Resistance Events. This timeline illustrates the chronological emergence of antibiotic-resistant bacteria alongside the introduction of corresponding antibiotics from 1940 to 2011. Each entry highlights significant discoveries, including the identification of resistant strains and the development of new antibiotics. The timeline underscores the ongoing challenge of antibiotic resistance in various bacterial species and the continual efforts in pharmaceutical advancements. This figure has been adapted from the article by Ventola, 2015.
Abbreviations: PDR = pan-drug-resistant; R = resistant; XDR = extensively drug-resistant.

1.2. STATISTICAL OVERVIEW OF ANTIMICROBIAL RESISTANCE

1.2.1. ANTIMICROBIAL RESISTANCE IN EUROPE AND EUROPEAN ECONOMIC AREA

Antimicrobial Resistance (AMR) is perhaps the greatest perennial threat to public health. It is estimated that there is a total of 671,689 infections with antibiotic-resistant bacteria in the Europe and European Economic Area (EEA) in 2015. These infections were associated with approximately 33,110 deaths attributable to antibiotic resistance. Furthermore, infections accounted for 874,541 Disability-adjusted life-years (DALYs), indicating a significant health burden. DALYs are a measure of overall disease burden, expressed as the number of years lost, due to early death or ill-health. Interestingly, a notable 63.5% of the infections were associated with healthcare settings, highlighting the importance of hospital-related interventions. These statistics underscore the substantial impact of antibiotic-resistant infections on public health in the Europe and EEA (Cassini *et al.*, 2019).

According to the World Health Organization, AMR still remains a major public health concern. Similar to the results above, in 2020 there were more than 670,000 infections annually in the EU/EEA, due to resistant bacteria to antibiotics. There are also geographical variations, with higher AMR percentages in the southern and eastern parts of Europe (World Health Organization, 2022). Compared to the data for 2021, it turns out that the situation of AMR varied widely depending on the bacterial species, the antimicrobial group and the geographical region. The total number of isolates reported was higher in 2021 than in 2020, with higher numbers of isolates reported across all pathogens (World Health Organization, 2023). This is why the European Union has established measures to combat microbial resistance, such as promoting the use of antibiotics only when needed and in the right way and promoting research into the development of new antibiotics (World Health Organization, 2023). Finally, the related cost to the health care systems of these countries is estimated to be around €1.1 billion (World Health Organization, 2022).

1.2.2. ANTIMICROBIAL RESISTANCE LANDSCAPE IN GREECE

The antimicrobial resistance landscape in Greece has also shown notable trends based on data from the aforementioned European Centre for Disease Prevention and Control (ECDC) reports. Greece continues to experience high levels of AMR, particularly concerning Gram-negative bacteria. In 2020, Greece had some of the highest rates of carbapenem resistance in *Klebsiella pneumoniae* (66.3%) and *Acinetobacter species* (94.6%) in Europe. Resistance to third-generation cephalosporins in *Escherichia coli* also remained significant at 21.9%.

The 2021 data reinforced these concerning trends, showing persistently high levels of resistance. Carbapenem resistance in *Klebsiella pneumoniae* increased critically at 73.7%, while *Acinetobacter species* continued to show alarming resistance rates at 96.9%. Resistance to third-generation cephalosporins in *Escherichia coli* slightly decreased at 21.7%. Moreover, Greece reported high levels of combined resistance in *Pseudomonas aeruginosa* and *Escherichia coli*, further complicating treatment options and highlighting the need for improved infection control and antimicrobial stewardship. These data emphasize the ongoing challenges Greece faces in tackling AMR, necessitating robust public health strategies and international cooperation to mitigate the impact of resistant infections on healthcare systems (World Health Organization, 2022; World Health Organization, 2023).

1.2.3. ANTIMICROBIAL RESISTANCE LANDSCAPE IN THE UNITED STATES

In the USA, over 2.8 million Multi-drug Resistance (MDR) bacterial infections occur each year, resulting in 35,000 deaths and \$20 billion in health-care costs (Strathdee *et al.*, 2020). The Center for Disease Control and Prevention (CDC) continues to monitor and combat AMR through various initiatives, including the National Healthcare Safety Network and the Antimicrobial Resistance Laboratory Network, which track resistance patterns and support rapid response efforts.

In 2013, AMR remained a significant public health concern in the United States. CDC reported that more than 2 million infections and at least 23,000 deaths annually were attributed to antibiotic-resistant bacteria. Key pathogens of concern included Carbapenem-resistant *Enterobacteriaceae*, drug-resistant *Neisseria gonorrhoeae*, methicillin-resistant *Staphylococcus aureus* and various resistant strains of *Clostridium difficile* (Centers for Disease Control and Prevention, 2013). However, 2019 data show some improvements, such as a 16% decrease in hospital-onset methicillin-resistant *Staphylococcus aureus* and a 3% decrease in hospital-onset *Clostridioides difficile* infections (Centers for Disease Control & Prevention, 2019). Challenges include resource diversion, antibiotic overuse, delays in legislation, drug shortages, and low public awareness. Opportunities lie in surveillance, stewardship, rapid diagnostics, a One Health approach, and incentivizing new antibiotics (Strathdee *et al.*, 2020).

1.2.4. ANTIMICROBIAL RESISTANCE AND COVID-19 RELATION

The impact of the COVID-19 pandemic on antimicrobial resistance has been significant. During the pandemic, patients with COVID-19 received antimicrobial treatment even when bacterial or fungal coinfections were less common (Hsu, 2020). Frequent use of antibiotics in patients with COVID-19 has the potential to worsen AMR, underscoring the urgent need for future research and surveillance at both patient and population levels (Langford *et al.*, 2023).

U.S. healthcare facilities, especially hospitals, saw more healthcare-associated, antimicrobial-resistant infections. Hospitals treated sicker patients who required prolonged use of medical devices like catheters and ventilators. Challenges such as shortages of personal protective equipment, staffing issues, and longer patient visits contributed to increased resistant infections (Centers for Disease Control and Prevention, 2020). Hospital-onset infections and deaths from resistant pathogens increased by at least 15% from 2019 to 2020. Data on the spread of antimicrobial resistant pathogens in communities were limited due to diverted public health personnel and reduced access to care and testing services during the pandemic (Centers for Disease Control and Prevention, 2020). Overall, the pandemic disrupted the detection and reporting of AMR data, leading to data gaps and hindering analysis. While nationwide investments in prevention had previously driven down deaths from antimicrobial resistant infections, the pandemic resulted in more resistant infections, increased antibiotic use, and less data and prevention actions (National Center for Emerging and Zoonotic Infectious Diseases, 2022). Properly resourcing public health and healthcare systems remains crucial to combat AMR effectively.

1.3. MODES OF ANTIBIOTICS

Research in the field of antibiotics and their mode of action is a promising topic in academia and industry. As biomolecules, they offer complex intellectual puzzles and surprises regarding their chemical makeup, biosynthetic pathways, evolutionary history, and biochemical actions (Brötz-Oesterhelt and Brunner, 2008). Reproducing these natural compounds synthetically in the laboratory poses significant difficulties due to their intricate structures and chirality. For instance, although the antibiotic penicillin was discovered in 1928, its complete

molecular structure remained a mystery until 1949 when Dorothy Crowfoot Hodgkin's X-ray crystallography studies elucidated it. This discovery was later confirmed through total synthesis efforts in 1959 (Hodgkin, 1949; Sheehan and Henery-Logan, 1959).

Antibiotics primarily exert their effects through various mechanisms, targeting essential processes in bacterial cells. One common mode of action is the inhibition of cell wall synthesis. Beta-lactam antibiotics, such as penicillin or cephalosporin, bind to and inactivate binding proteins related to these bacteria that are crucial for cell wall construction. Without a functional cell wall, bacteria are unable to maintain their shape and integrity, leading to cell lysis and death (Spratt, 1975; Tipper and Strominger, 1965). Another vital mechanism involves disruption of protein synthesis. Aminoglycosides, for example, bind to the 30S bacterial ribosomal subunit, causing errors in mRNA translation into proteins. This results in the production of dysfunctional proteins, ultimately leading to bacterial cell death (Davies and Davis, 1968).

Moreover, antibiotics can target bacterial DNA replication and repair mechanisms. Quinolones, such as ciprofloxacin, inhibit bacterial DNA gyrase and topoisomerase IV, enzymes critical for DNA supercoiling and replication. This inhibition prevents proper DNA replication and transcription, causing lethal genetic damage (Drlica and Zhao, 1997). Additionally, antibiotics like rifamycins (Figure 1.2) interfere with RNA synthesis by binding to bacterial RNA polymerase, thereby blocking the transcription process (Wehrli and Staehelin, 1971). These various modes of action not only highlight the versatility of antibiotics, but also underscore the necessity of precise targeting to avoid off-target effects and resistance development in bacterial populations.

In addition to these primary mechanisms, antibiotics can also disrupt bacterial metabolism and energy production. For instance, sulfonamides (Figure 1.2), such as sulfamethoxazole, inhibit the synthesis of folic acid by competitively binding to the enzyme dihydropteroate synthase, an essential step in the production of tetrahydrofolate. This inhibition halts the synthesis of nucleotides and amino acids, ultimately leading to bacterial cell death (Brown, 1962). Another example is metronidazole, which targets anaerobic bacteria and protozoa by entering the cells and undergoing reduction to form toxic intermediates that damage DNA and other vital cellular components (Edwards, 1993). These additional modes of action further illustrate the diverse strategies employed by antibiotics to combat bacterial infections and highlight the ongoing need for research to discover novel targets and mechanisms to counteract antibiotic resistance.

1.4. GENETICS OF RESISTANCE

The genetics of antimicrobial resistance is a complex and rapidly evolving field that underpins the mechanisms by which bacteria develop and propagate resistance to antibiotics. One key aspect involves Horizontal Gene Transfer (HGT) of resistance genes, which can occur through mechanisms such as conjugation, transformation, and transduction. Plasmids, transposons, and integrons play a critical role in this process by facilitating the transfer of genetic material between bacteria, often across different species. These mobile genetic elements can carry multiple resistance genes, leading to the development of multidrug-resistant strains. Whole Genome Sequencing (WGS) and other genomic tools have been pivotal in identifying and characterizing these elements, helping researchers understand the dissemination patterns of resistance genes across different environments and host populations (Djordjevic *et al.*, 2023; Julian Davies and D. Davies, 2010).

Moreover, the evolution of antimicrobial resistance is driven by genetic mutations that alter target sites, reduce drug uptake, increase efflux, or inactivate antibiotics. Studies have shown that the selective pressure exerted by the misuse and overuse of antibiotics accelerates the emergence of these mutations. Surveillance and genomic analysis have revealed the spread of resistant strains, particularly in healthcare settings where antibiotic use is prevalent. The ESKAPEE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter species*, and *Escherichia coli*)

Mechanisms of Antibacterial Drugs

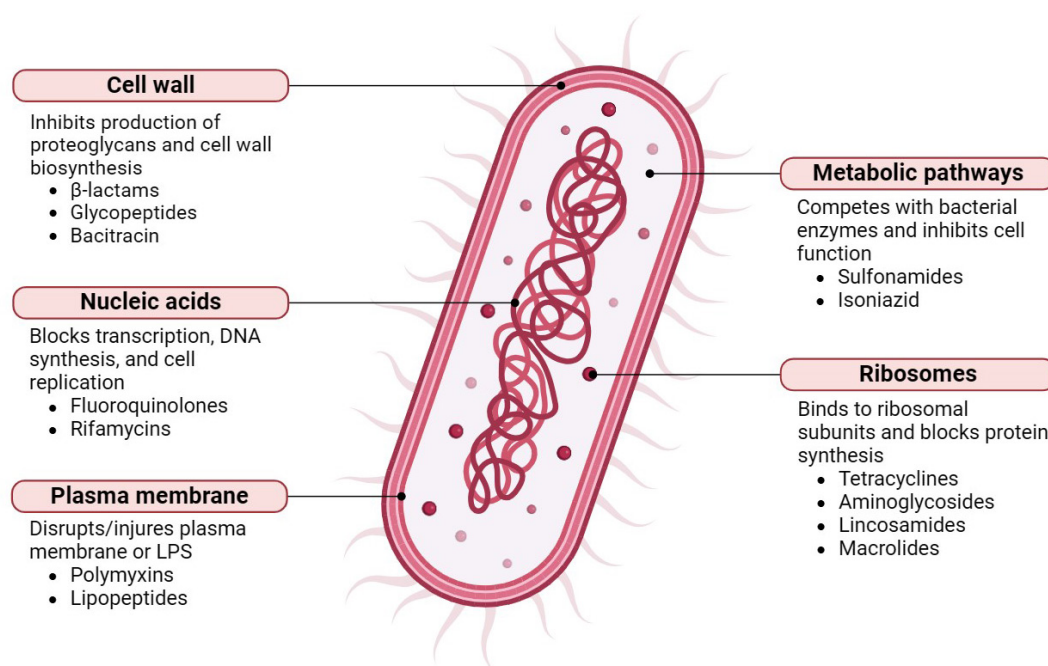


Figure 1.2 | The image illustrates the various mechanisms of action of antibacterial drugs. These mechanisms include inhibition of cell wall biosynthesis by β -lactams, glycopeptides, and bacitracin; blockage of transcription, DNA synthesis, and cell replication by fluoroquinolones and rifamycins; disruption of the plasma membrane by polymyxins and lipopeptides; competition with bacterial enzymes by sulfonamides and isoniazid; and inhibition of protein synthesis by tetracyclines, aminoglycosides, lincosamides, and macrolides.

exemplify critical threats due to their multidrug-resistant nature. Addressing AMR requires an integrated One Health approach, considering the interconnectedness of human, animal, and environmental health, and necessitates collaborative efforts across disciplines to monitor, understand, and combat the genetic mechanisms driving resistance (Woldegeorgis *et al.*, 2023).

However, the persistence and proliferation of AMR are facilitated not only by genetic mutations and HGT, but also by the adaptive capacity of bacterial communities within various ecological niches. Environmental reservoirs, such as soil and water, play significant roles in maintaining and spreading Antimicrobial Resistant Genes (ARGs). These environments often contain diverse microbial populations that exchange genetic material, including resistance determinants, through mechanisms such as conjugation and transformation. Studies have indicated that agricultural practices, particularly the use of antibiotics in livestock, contribute significantly to the reservoir of resistance genes, which can then be transferred to human pathogens (Djordjevic *et al.*, 2023).

Additionally, the complexity of the microbial resistome—the collection of all resistance genes in both pathogenic and non-pathogenic bacteria—underscores the challenge in managing AMR. Advanced genomic tools, such as metagenomics and resistome profiling, have revealed the vast diversity and mobility of resistance genes across different ecosystems. These tools have also helped identify latent resistance genes that could become significant in clinical settings. Effective AMR management thus requires robust genomic surveillance and a deeper understanding of the ecological dynamics of resistance gene spread. Interventions must be multi-faceted, encompassing prudent antibiotic use, stringent infection control practices, and international collaboration to track and mitigate the global spread of resistance genes. In conclusion, the genetics of antimicrobial resistance is a critical area of research that combines molecular biology, genomics, and epidemiology. Understanding the genetic basis and transmission dynamics of resistance genes is essential for developing strategies to combat AMR (Djordjevic *et al.*, 2023; Woldegeorgis *et al.*, 2023).

1.5. ECOLOGICAL ROLES OF ANTIMICROBIAL RESISTANCE

Antimicrobial Resistance (AMR) significantly impacts ecological dynamics by altering microbial community structures and interactions. Antibiotics, while revolutionary in the treatment of infections, have also led to the rise of resistant strains in various environments. The overuse and misuse of antibiotics in clinical, agricultural and other settings have created strong selection pressures, allowing resistant microbes to thrive. These resistant strains can transfer their resistance genes to other bacteria through HGT, further spreading resistance. This phenomenon is not limited to human pathogens but extends to environmental microbiomes, where resistance genes can be found in natural microbial populations. The presence of these genes in the environment suggests that resistance is a natural phenomenon, exacerbated by human activities (Julian Davies and D. Davies, 2010).

Moreover, the ecological impact of antibiotic resistance is profound. Antibiotics and resistant bacteria can enter ecosystems through various pathways, including wastewater, agricultural runoff, and improper disposal of medications. This contamination affects microbial communities in soil, water, and other habitats, potentially disrupting ecological balances. Resistant bacteria can outcompete susceptible strains, leading to shifts in microbial diversity and function. Additionally, the spread of resistance genes in the environment poses a risk to human and animal health, as these genes can be acquired by pathogenic bacteria, making infections harder to treat. Understanding the ecological roles of antibiotic resistance is crucial for developing strategies to mitigate its spread and impact (Julian Davies and D. Davies, 2010). This natural exchange of genetic material enables microbial communities to rapidly adapt to environmental stresses, including antibiotic exposure. The dissemination of ARGs among diverse microbial taxa can lead to the emergence of multidrug-resistant strains, posing challenges to public health and ecosystem stability (Martinez, 2009).

The role of antibiotics in natural environments extends beyond their therapeutic use, serving as signaling molecules and mediators of microbial interactions. In soil and aquatic habitats, for instance, antibiotics produced by microorganisms can inhibit competitors, regulate population dynamics, and influence community structure. These compounds can act as cues in microbial communication, affecting processes such as biofilm formation, spore germination, and nutrient cycling. The production of antibiotics by environmental microbes suggests that these compounds have evolved primarily as ecological tools rather than just defense mechanisms against pathogens. This ecological perspective highlights the complexity of microbial ecosystems and the multifaceted roles that antibiotics play within them (Martinez, 2009).

1.6. NOSOCOMIAL INFECTIONS

Nosocomial infections, or Hospital-acquired Infections (HAIs), pose a significant challenge to healthcare systems worldwide, particularly when associated with ESKAPEE pathogens. The acronym ESKAPEE refers to a group of bacteria known for their ability to "escape" the effects of antibiotics and cause severe infections. These pathogens include *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter* species and *Escherichia coli*. These bacteria are notorious for their multidrug resistance, which complicates treatment options and leads to increased morbidity and mortality among hospitalized patients (Boucher *et al.*, 2009; Santajit and Indrawattana, 2016).

The rise of AMR in ESKAPEE pathogens exacerbates the threat of nosocomial infections. AMR occurs when bacteria develop mechanisms to withstand drugs designed to kill them. This resistance can be intrinsic or acquired through genetic mutations and horizontal gene transfer. ESKAPEE pathogens have demonstrated a remarkable ability to develop and share resistance genes, rendering many commonly used antibiotics ineffective (Munita and Arias, 2016). For instance, methicillin-resistant *Staphylococcus aureus* and carbapenem resistant *Enterobacteriaceae* are prominent examples of how resistance can severely limit treatment options, leading

to prolonged hospital stays, higher healthcare costs, and increased mortality rates (Tacconelli *et al.*, 2018). Ultimately, addressing the challenge of ESKAPEE-related nosocomial infections demands coordinated efforts from healthcare providers, researchers, policymakers, and the public (World Health Organization, 2014).

The genetic mechanisms underlying antimicrobial resistance in ESKAPEE pathogens are diverse and complex. These bacteria possess various resistance determinants, including beta-lactamases, which hydrolyze the beta-lactam ring of antibiotics such as penicillins and cephalosporins, rendering them ineffective. For example, *Klebsiella pneumoniae* carbapenemases (KPCs) and New Delhi metallo-beta-lactamases (NDM-1) are enzymes that confer resistance to a wide range of beta-lactam antibiotics, including last resort carbapenems (Munita and Arias, 2016). Additionally, efflux pumps, such as those found in *Pseudomonas aeruginosa*, actively expel antibiotics from bacterial cells, reducing intracellular drug concentrations and efficacy (Poole, 2001). Mutations in target sites, such as alterations in penicillin-binding proteins in *Staphylococcus aureus*, also contribute to resistance by decreasing antibiotic binding affinity (Chambers and DeLeo, 2009). These mechanisms, combined with horizontal gene transfer of resistance genes through plasmids, transposons, and integrons, enable ESKAPEE pathogens to rapidly acquire and disseminate resistance traits, complicating infection control and treatment strategies (Partridge *et al.*, 2018).

1.7. ANTIMICROBIAL SUSCEPTIBILITY TESTING (AST)

Antimicrobial Susceptibility Testing (AST) is a critical laboratory procedure used to determine the sensitivity of bacterial pathogens to various antimicrobial agents. By assessing whether bacteria are susceptible, intermediate, or resistant to specific antibiotics, AST plays a major role in guiding effective treatment plans for infections. The process involves exposing bacteria isolated from clinical samples to antimicrobial agents and observing the extent of inhibition of bacterial growth. This is typically done through methods such as disk diffusion, broth dilution, and automated systems (Reller *et al.*, 2009). The results from AST inform clinicians about the most appropriate antimicrobial therapy, thereby enhancing patient outcomes and helping to combat the rise of antimicrobial resistance (Turnidge and Paterson, 2007).

A key measure obtained from AST is the Minimum Inhibitory Concentration (MIC), which represents the lowest concentration of an antimicrobial agent that effectively inhibits visible bacterial growth. MIC values are critical for determining the susceptibility of a microorganism to an antibiotic. These values are obtained through standardized testing methods such as broth microdilution, which involves exposing bacteria to a range of antibiotic concentrations in a controlled environment (Andrews, 2001). MIC values provide a quantitative metric that helps in categorizing bacteria as susceptible, intermediate, or resistant based on established breakpoints set by organizations like the Clinical and Laboratory Standards Institute (CLSI) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (Wiegand *et al.*, 2008).

Antibiograms are cumulative reports generated from AST results that summarize the susceptibility profiles of bacterial isolates within a particular institution or region over a specified period. These reports are invaluable for both clinicians and epidemiologists as they provide insights into local resistance patterns and trends. Antibiograms assist in empirical therapy decisions, particularly in critical cases where timely treatment is essential, and they also aid in formulating antimicrobial stewardship policies (CLSI, 2020). By identifying prevalent resistance patterns, antibiograms help in monitoring the effectiveness of infection control interventions and in adjusting antibiotic use protocols to minimize the development and spread of resistant strains (CLSI, 2020).

In summary, AST, MIC values, and antibiograms collectively play a crucial role in the management of bacterial infections and the fight against antimicrobial resistance. AST provides essential data on the efficacy of antibiotics, MIC values offer precise susceptibility metrics, and antibiograms enable the analysis of resistance

trends and inform public health strategies. Together, they ensure the judicious use of antimicrobial agents, ultimately contributing to better patient care and the mitigation of resistance issues.

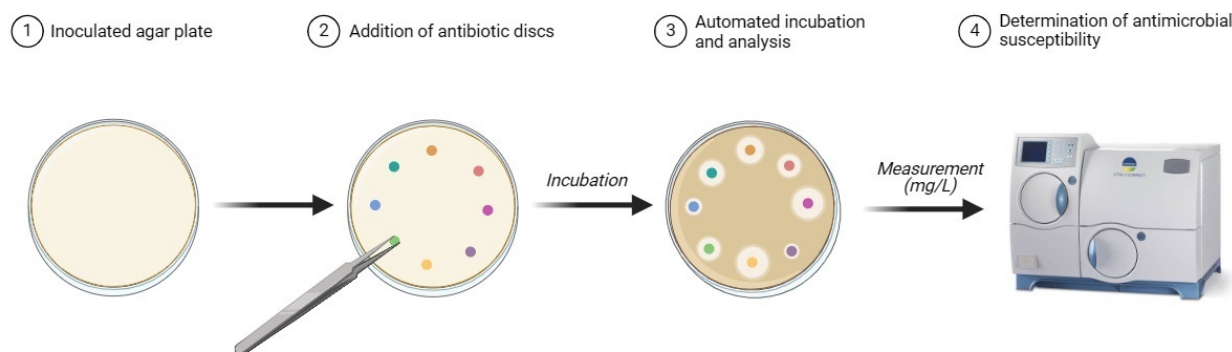


Figure 1.3 | Antimicrobial Susceptibility Testing (AST) Process. The process involves inoculating an agar plate with bacteria, placing antibiotic discs on the surface, and incubating the plate. Zones of inhibition, where bacterial growth is prevented around the discs, are measured to determine the bacteria's susceptibility to the antibiotics.

1.8. ANTIMICROBIAL RESISTANCE DATABASES

Antimicrobial resistance databases are vital bioinformatics tools designed to track, analyze, and manage the growing threat of drug-resistant pathogens. These databases compile comprehensive information on resistance genes, mutations, and mechanisms that bacteria, viruses, fungi, and parasites employ to evade antimicrobial treatments. By integrating genomic, phenotypic and epidemiological data, AMR databases enable researchers and healthcare professionals to identify emerging resistance patterns, predict outbreaks, and develop targeted strategies for treatment and prevention. Tools such as Comprehensive Antibiotic Resistance Database (CARD), ResFinder, and ARG-ANNOT facilitate the annotation and interpretation of resistance determinants in sequenced microbial genomes, supporting the development of new diagnostics, therapeutics, and public health policies aimed at combating antimicrobial resistance globally. In this work, we will focus on using the following databases to derive genotypic and phenotypic antimicrobial resistance data and analyze them in the methodology section.

1.8.1. NATIONAL DATABASE OF ANTIBIOTIC RESISTANT ORGANISMS

The National Database of Antibiotic Resistant Organisms (NDARO) is a centralized resource created by the National Center for Biotechnology Information (NCBI) to combat antimicrobial resistance. Established as part of the 2015 National Action Plan for Combating Antibiotic Resistant Bacteria, NDARO serves as a collaborative, cross-agency platform that provides researchers and public health officials with access to comprehensive AMR data. The database facilitates real-time surveillance of pathogenic organisms by providing curated information on AMR genes and a collection of genetic and antibiotic susceptibility data. This resource is instrumental in tracking the evolution and spread of antibiotic resistance, helping to inform and enhance strategies to mitigate this global health threat (NCBI, 2023).

NDARO's contributions to AMR research include the development of tools such as AMRFinderPlus, which identifies AMR genes in bacterial genomes, and the Isolate Browser, which allows users to locate bacterial genomes with specific resistance traits. The database also plays a critical role in standardizing and disseminating AMR-related data, ensuring that vital information is accessible for effective decision-making in both clinical

and research settings. By partnering with organizations like the CDC, FDA, USDA and WHO, NDARO helps coordinate efforts to address the pressing issue of antibiotic resistance on multiple fronts (NCBI, 2023).

To date, NDARO database houses extensive data on over 1,800,000 pathogen isolates, including detailed information on their genetic makeup and antibiotic resistance profiles. This repository features data on multiple drug-resistant bacteria, encompassing both healthcare-associated pathogens and community-acquired strains. The NDARO is continually updated through contributions from various health facilities, research institutions, and public health laboratories, making it a critical resource for monitoring and combating antimicrobial resistance. This robust database supports the identification of resistance trends and aids in the development of targeted strategies to mitigate the spread of resistant organisms (NCBI, 2023).

Link: [NDARO Database](#)

1.8.2. BACTERIAL AND VIRAL BIOINFORMATICS RESOURCE CENTER

The Bacterial and Viral Bioinformatics Resource Center (BV-BRC) is an extensive information system that supports research on bacterial and viral infectious diseases. It integrates data and tools from the PATRIC, IRD, and ViPR resources, providing researchers with a unified platform to access genomic data, protein structures, AMR genotype data and more. The BV-BRC hosts over a million viral genomes and hundreds of thousands of bacterial genomes, offering a variety of analytical tools such as genome assembly, annotation, phylogenetic trees, and metagenomic analysis. This integration enhances the efficiency and scalability of bioinformatics research, facilitating the exploration of pathogens at a large scale (Olson *et al.*, 2022)

The BV-BRC database is highly scalable, managing billions of records across various data types such as genomes, transcriptomics, protein functions, and epidemiological data. As of August 2022, it operates on a distributed architecture with high-performance computing capabilities. The workspace environment supports bulk data upload, storage, and sharing, and has been used successfully for over a decade. The comprehensive application framework enables asynchronous high-throughput analyses, integrating tools for quick searches and multi-day computational tasks, ensuring robust and efficient research support.

In terms of statistics, BV-BRC encompasses more than 2,100,000 bacterial genomes with annotated AMR genes. For example, the database includes information on common pathogens such as *Staphylococcus aureus*, *Escherichia coli* and *Pseudomonas aeruginosa*. It leverages machine learning algorithms to predict AMR phenotypes and includes advanced visualization tools that allow researchers to explore AMR gene distributions and related genomic features, enhancing the ability to track resistance trends.

Link: [BV-BRC Database](#)

1.8.3. NATIONAL ANTIMICROBIAL RESISTANCE MONITORING SYSTEM

The National Antimicrobial Resistance Monitoring System (NARMS), established in 1996, is a collaborative public health surveillance program involving the CDC, FDA, and USDA. Its primary focus is to monitor antimicrobial resistance in enteric bacteria from humans, retail meats, and food animals in the United States. Over the past two decades, NARMS has significantly contributed to public health by tracking changes in bacterial resistance, understanding how resistance spreads, and distinguishing between resistant and susceptible infections. This integrated surveillance has provided critical data to inform antimicrobial policies, guide treatment recommendations, and enhance food safety measures. NARMS data are essential for ensuring safe antimicrobial drug use, and addressing broader food safety priorities. The program continues to adapt to

emerging resistance threats and new technologies, such as whole genome sequencing (Karp *et al.*, 2017).

The NARMS database is tracking resistance trends among enteric bacteria from human, animal, and retail meat sources. According to the latest data, NARMS has collected over 180,000 bacterial isolates from humans, more than 110,000 from retail meats, and nearly 220,000 from food animals. This surveillance has shown a rise in multidrug-resistant bacteria, particularly among *Salmonella* and *Campylobacter* species, which are common causes of foodborne illnesses. The database reveals significant geographical and temporal variations in resistance patterns, helping to inform targeted public health interventions and antibiotic stewardship efforts across the United States.

Link: [NARMS Database](#)

1.9. MACHINE LEARNING ALGORITHMS

Machine learning (ML) has become an increasingly powerful tool in predicting AMR, offering significant advancements over traditional methods. By leveraging vast amounts of genomic, phenotypic and clinical data, ML algorithms can identify complex patterns and relationships indicative of resistance that may be imperceptible to human analysts. Recent advances in machine learning, such as supervised learning and deep learning are employed to develop predictive models that can accurately forecast resistance profiles based on genetic sequences or other biological markers. These models can assist in identifying novel resistance mechanisms, optimizing treatment regimens, and enhancing diagnostic accuracy. Notable applications include the use of ML to predict resistance from Whole Genome Sequencing (WGS) data, identify potential new antimicrobial targets, and forecast the spread of resistant strains. By improving the speed and accuracy of resistance detection, machine learning holds great promise in the global effort to address antimicrobial resistance and support clinical microbiology practices.

An early study by Nguyen *et al.*, 2018 explored the development of an in silico Minimum Inhibitory Concentration (MIC) panel test for *Klebsiella pneumoniae*. Using genomic data, the study applied random forest, Support Vector Machine (SVM) and XGBoost algorithms to predict MIC values, achieving an accuracy of 90% for the random forest model and 88% for the Support Vector Machine model, showcasing the feasibility of such approaches in simulating conventional laboratory tests for AMR determination. Similarly, Kim *et al.*, 2022 reviewed current practices and limitations in applying ML to AMR prediction, emphasizing the need for robust validation and integration into clinical workflows to ensure reliable and actionable outcomes. The study highlighted that ensemble methods, such as gradient boosting machines and random forests, typically achieve accuracies ranging from >80% to 95% depending on the data and specific pathogen considered.

ML techniques have also been tailored to identify AMR in the Intensive Care Unit (ICU) settings, focusing on the rapid and accurate detection of resistant pathogens. Martínez-Agüero *et al.*, 2019 highlighted the application of ML models in ICU environments, underscoring their potential to enhance decision-making and patient outcomes by providing timely resistance profiles. The study utilized deep learning and random forest algorithms, achieving an overall accuracy of around 85% for predicting AMR profiles. Furthermore, Macesic *et al.*, 2017 discussed novel bioinformatics approaches for combating AMR, detailing how ML can integrate vast genomic data to predict resistance mechanisms and guide effective treatments. This study also employed random forest and SVMs, reporting prediction accuracies between 80% and 90% across different datasets and bacterial species.

The predictive power of has also been demonstrated in specific pathogens. Khaledi *et al.*, 2020 used ML-enabled molecular diagnostics to predict resistance in *Pseudomonas aeruginosa*, achieving a high accuracy of 92% with ensemble algorithms and highlighting the importance of incorporating molecular data, like gene

expression and short nucleotide polymorphisms, for precise predictions. Similarly, a study by Moradigaravand *et al.*, 2018 on *Escherichia coli* utilized large-scale pan-genome data to predict antibiotic resistance, illustrating the scalability of ML approaches across diverse bacterial populations. The study employed gradient boosting machines and SVMs, achieving average accuracy of 91% (range 0.81–0.97).

Recent advancements have focused on leveraging Whole Genome Sequencing (WGS) data for AMR prediction. Ren *et al.*, 2021 and colleagues demonstrated that integrating WGS data with ML algorithms such as gradient boosting and Convolutional Neural Networks (CNNs) could accurately predict resistance profiles, with achieved accuracies of up to 94% depending on the bacterial species and resistance determinants considered. Additionally, Van Camp *et al.*, 2020 showed that ML models could predict AMR in gram-negative bacteria from WGS data, with accuracies ranging from 80% to 90% XGBoost algorithm. Similarly, a study on nontyphoidal Salmonella by Nguyen *et al.*, 2019 utilized ML to predict antimicrobial MICs and associated genomic features, achieving prediction accuracies of 88% with a XGBoost model.

In summary, the integration of machine learning with genomic data presents a promising avenue for improving AMR prediction. While current practices have shown significant potential, challenges such as data standardization, model validation, and clinical integration remain. Continued advancements in bioinformatics and the increasing availability of genomic data are likely to enhance the accuracy and utility of ML models in predicting antimicrobial resistance. Across these studies, the format of the data varied from Whole Genome Sequencing data to pan-genome data, with predictions typically covering a range of 4 to 11 different antibiotics depending on the pathogen and the specific focus of the study. Only two studies contained prediction in 15 antibiotics and above (up to 20), indicating that constructing a comprehensive dataset with many genomes and antibiotics is imperative to improve the performance of algorithms.

2 | METHODS

In the present work, we will use Machine learning to predict Antimicrobial Resistance from Whole Genome Sequencing data, derived from public databases. We will focus only on seven common nosocomial organisms, Gram-positive and Gram-negative bacteria, that can cause sepsis (ESKAPEE pathogens): *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter species* and *Escherichia coli*. These specific pathogens were chosen because, in addition to clinical relevance, they are the most represented in terms of AMR-annotated samples available at NCBI.

2.1. PUBLIC DATA COLLECTION

We obtained the requested information from the Isolates Browser on the NCBI Pathogen Detection website (NDARO database) (NCBI, 2023), applying filters for genus (restricted to ESKAPEE pathogens genera) and AST phenotypes (Resistant, Susceptible, Intermediate, and Not detected). This data set comprised 13,827 isolates, corresponding to 12,865 unique assemblies, as some isolates lacked an NCBI assembly ID. The table includes details such as organism, strain, NCBI BioProject, BioSample and Assembly IDs, contig information, etc. BioSample IDs were utilized to retrieve antibiogram tables using the Entrez system (Schuler *et al.*, 1996). Each antibiogram includes information on antibiotics, measurement signs, values and units, laboratory methods, etc. Hence, MIC values were extracted from the BioSample IDs.

We concatenated all the antibiograms into a single comprehensive data frame. Initially, we retained only unique biosamples, removing duplicates using an R script. Subsequently, assemblies reported with antibiograms' measurement unit in millimeters (638 assemblies) were excluded. Using a custom Snakemake pipeline, (Mölder *et al.*, 2021) we have downloaded assemblies, based on NCBI Assembly IDs, we have downloaded biosamples without assembly IDs (only SRA IDs), and converted them into assemblies using the SPAdes tool (Bankevich *et al.*, 2012), for 962 assemblies, with parameters `--only-assembler`, utilizing the Entrez system. Of the 962 IDs, 937 were successfully retrieved, with the remainder excluded due to conversion error, because of unbalanced paired-end read data. Finally, we achieved a total of 13,159 unique assemblies with a corresponding number of antibiograms. Also, we have splitted antibiograms per antibiotic and we have created a Comma-separated Values (CSV) file containing each antibiogram, and a CSV file with NDARO strain metadata.

We obtained data from the NARMS website, applying the following filters: Test Method (WGS), Bacteria (*E. coli* O157), Antibiotic (All, 18), Date Range (2016 to 2024), and States (All). This initial dataset comprised 6,616 isolates in xlsx format. After filtering out rows with missing antibiogram data using an R script, we retained 477 isolates, corresponding to 477 unique BioSample IDs. The data were transformed to ensure consistency with antibiograms from other databases, which involved removing columns such as Year of Isolation, Samples Region Name, and source. The original dataset contained two ID columns: NCBI Accession ID and WGS ID. We combined these into a single column, using the NCBI Accession ID where the WGS ID was missing, and removed rows with missing ID information. Subsequently, we split this table into individual data frames for each antibiotic and data from these IDs were downloaded using a Snakemake pipeline and the Entrez system. The final dataset includes 477 unique assemblies, 14 CSV files with split antibiograms per antibiotic, a comprehensive CSV file containing all antibiograms, and a CSV file with NARMS strain metadata.

In the next step, we have downloaded the complete AMR table from the BV-BRC database, applying filters for Resistant Phenotype (Resistant, Susceptible, Intermediate, and Not detected) and Measurement Unit (mg/L). This initial dataset contained 174,256 records. We further filter these isolates, retaining only those belonging to the genera *Klebsiella*, *Enterococcus*, *Acinetobacter*, *Pseudomonas*, *Staphylococcus*, *Enterobacter*, and *Escherichia* (ESKAPEE pathogens), using an R script, which results in 103,012 records. Furthermore, we preprocessed the filtered data by converting the measurement values to numeric format, selecting specific

columns of interest, and transforming the Patric Genome IDs into NCBI Assembly or SRA IDs based on the BV-BRC genome table. For some SRA IDs, we converted them to SRA accession numbers using the Entrez system to facilitate the downloading of sequence files. This preprocessing step was performed using R, followed by data downloading via our Snakemake pipeline.

During the processing of SRA IDs, we also encountered issues with the SPAdes tool for assembly creation, due to some SRA files containing an unequal number of reads, likely due to sequencing technical errors. Consequently, we excluded 3,068 records (161 unique assemblies) from our analysis, resulting in 6,249 unique assemblies. The final data frame contained 99,944 records. We finally obtained 6,249 unique assemblies, 85 CSV files with separated antibiograms per antibiotic, a comprehensive CSV file containing all antibiograms, and a CSV file with strain metadata.

Finally, from the three databases used, we obtained a total of 19,885 records. This dataset can be validated as it comprises 13,159 unique strains from NDARO, 477 from CDC NARMS, and 6,249 from BV-BRC, with a total of 753 overlapping strains between the NDARO and BV-BRC databases. Thus, we have **19,132 unique final assemblies**. Duplicate IDs were filtered, where we retained rows from NDARO due to their more comprehensive metadata.

2.2. SNAKEMAKE PIPELINE FOR GENOMIC DATA ANALYSIS

At this stage we deem it necessary to mention a few things about the Snakemake software that we used extensively in our analysis, before proceeding further to the methodology. Snakemake software (Mölder *et al.*, 2021) is a versatile and efficient workflow management system, including a specification language on top of Python, designed for bioinformatics and other scientific data analysis. It streamlines the process of creating reproducible and scalable workflows, enabling researchers to manage complex data processing tasks with ease. In this study, we leveraged Snakemake to automate the steps involved in downloading, processing, and analyzing genomic data from public databases, ensuring consistency and accuracy throughout the entire workflow.

One of the key advantages of using Snakemake is its ability to automate the execution of various data processing steps based on a predefined workflow. In our research, Snakemake was used to download genomic assemblies and associated metadata from the NCBI Pathogen Detection database (NDARO), NARMS, and BV-BRC databases. By defining rules and dependencies in a Snakefile, we ensured that each step of the pipeline was executed in the correct order, from initial data retrieval to final data processing and quality control. Snakemake seamlessly integrates with a wide range of bioinformatics tools, allowing us to perform complex genomic analyses efficiently. For instance, we used SPAdes for genome assembly from raw sequencing reads. Snakemake facilitated the conversion of SRA IDs to assembly data, ensuring that all necessary steps were performed automatically. This integration significantly reduced the manual effort required to handle large datasets and complex bioinformatics workflows.

By automating these steps, Snakemake ensured that the data processing was consistent, reproducible, and free from human errors. This level of automation is crucial for handling large-scale genomic datasets and maintaining the integrity of the analysis. Snakemake's scalability allows us to run workflows on various computing environments, from local machines to high-performance computing clusters. In our study, the pipeline was executed on a high-performance server, to handle the extensive computational requirements of processing thousands of genomic assemblies. Furthermore, Snakemake enhances reproducibility by documenting the entire workflow, including software versions, parameters, and dependencies. This ensures that the analysis can be replicated by other researchers, promoting transparency and reliability in scientific research.

Finally, Snakemake optimizes resource allocation by parallelizing tasks and managing dependencies effectively.

This capability is particularly important when dealing with large datasets, as it minimizes computational time and maximizes resource utilization. In our analysis, Snakemake efficiently handled the downloading and processing of over 19,000 genomic assemblies, significantly reducing the overall time required to complete the workflow.

2.3. ASSEMBLIES QUALITY CONTROL

Assemblies Quality Control is a critical step in the genomic data analysis pipeline, ensuring that the assembled sequences are of high quality and suitable for downstream applications. One of the key metrics used for assessing the quality of genome assemblies is the N50 value. The N50 value is defined as the length of the shortest contig at which the sum of the lengths of all contigs of that length or longer contains at least 50% of the total assembly length. It serves as an indicator of the contiguity and completeness of the assembly; higher N50 values typically signify better assembly quality, as they reflect longer contiguous sequences. During the quality control process, assemblies with low N50 values may be flagged for further scrutiny or improvement, as they may indicate fragmentation or gaps in the assembly, or they can be completely removed from the analysis. By analyzing the N50 values, we can identify potential issues, and make informed decisions to enhance the overall quality of the genomic data. This metric, alongside other quality indicators, forms the foundation of a robust quality control framework, ensuring the reliability and accuracy of genomic assemblies used in subsequent analyses.

For the NDARO database, N50 values were directly obtained from the NDARO initial table. However, for records from the other two databases (CDC NARMS and BV-BRC), such information was not available. Therefore, we manually calculated the N50 values using the `Biostrings::N50` function (Pagès *et al.*, 2024). These calculated N50 values were then incorporated into the strain metadata file as an additional column, ensuring a consistent and comprehensive quality metric across all datasets.

Subsequently, we conducted an exploratory analysis to identify strains with low-quality assemblies. We plotted the mean N50 values across different genera and observed that *Enterobacter* had the highest mean N50 value, while *Escherichia coli* had the lowest, despite being the most abundant genus with 9,485 strains. Additionally, a box plot (Figure 2.1) was created to examine the variation in the number of outliers among different genera. Using an R script, we generated density plots for individual genera and determined appropriate N50 cutoff values. A general cutoff value of 50,000 was established, with specific cutoffs of 20,000 for *Enterobacter* and *Staphylococcus*. Applying these cutoff values resulted in the removal of 587 records from the initial set of 19,132 unique assemblies, leaving us with a final total of 18,545 unique strains. The final bar plot of the filtered and unfiltered dataset is provided below, as well as the density plot of the N50 values (Figure 2.1).

2.4. ANTIBIOGRAM TABLES

After quality control of assemblies, we created the final table of antibiograms by concatenating the data frames from the three databases, corresponding to 19,885 assemblies. However, we removed overlapping assemblies between the NDARO and BV-BRC databases, and assemblies rejected in the quality control of the previous step and the final concatenated data frame corresponded to 18,545 unique assemblies. A new column indicating the database source of each antibiogram was also added. To preprocess the data, antibiotic names were standardized to ensure uniformity, MIC signs were formatted consistently and antibiotic names corresponding to the same antibiotic were grouped. Standard typing values were also standardized to ensure uniformity, and signs and values were converted to a numeric format. The antibiograms table includes the columns shown in Table 2.1.

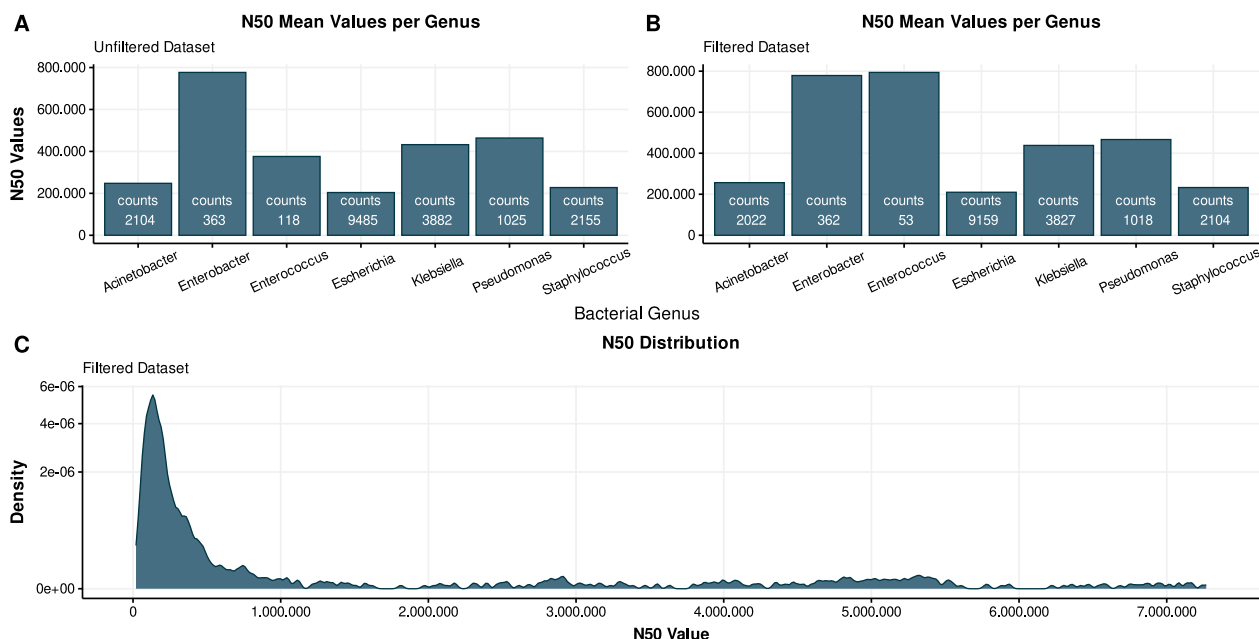


Figure 2.1 | Quality Control Analysis of Genomic Assemblies by Bacterial Genus. (A) *N50 Mean Values per Genus for Unfiltered Dataset:* The bar graph shows the mean N50 values for different bacterial genera. The counts of each genus are annotated within the bars, indicating the number of assemblies per genus in the unfiltered dataset. (B) *N50 Mean Values per Genus for Filtered Dataset.* (C) *N50 Value Distribution for Filtered Dataset:* The density plot depicts the distribution of N50 values across the filtered dataset, indicating the range and frequency of N50 values among the assembled genomes. The distribution highlights the variations in assembly quality across different bacterial genera.

Column	Description
antibiotic	Antibiotic names
sign	Measurement Sign (<, >, ≤, ≥, ==)
value	MIC value
measurement	Measurement Unit (mg/L)
phenotype	R: resistant, S: susceptible, I: intermediate, N: not detected
organism	Strain of organism
platform	Measurement Platform (e.g., Phoenix, Vitek, etc)
standard	Resistant/Susceptible/Intermediate Standard
id	ID of the database (SRA, Assembly, BioSample)

Table 2.1 | Description of Columns in the Antibigram Data Frame. This table lists and describes the columns included in the final combined antibigram data frame. Each column captures specific attributes of the antibigram data, crucial for subsequent analysis and interpretation.

2.5. ANTIBIOGRAMS QUALITY CONTROL

Quality control in antibiograms involves validation processes to ensure the accuracy and reliability of the susceptibility results. Key to this process are the known breakpoints, which are specific Minimum Inhibitory Concentration (MIC) thresholds that classify microorganisms as resistant, intermediate, or susceptible to particular antibiotics. Quality control procedures include the check of consistency between the reported phenotype and the reported MIC value. If the reported MIC values fall within the expected range of the reported phenotype, the strain is considered reliable. To identify erroneous MIC values, the reported MIC results are compared against the established breakpoints. Any discrepancies may indicate a problem with the reporting, testing procedure, equipment, or reagent quality and that's why the strain was deleted for the specific antibiotic.

For the scope of this thesis, we have conducted quality control on the Gentamicin antibiogram. Gentamicin is a broad-spectrum aminoglycoside antibiotic commonly used to treat various bacterial infections, particularly

those caused by Gram-negative bacteria (Narayanan and Lewis, 2023). It functions by binding to the 30S bacterial ribosomal subunit, leading to inhibition of protein synthesis, which ultimately results in bacterial cell death. Gentamicin is often administered intravenously or intramuscularly and is effective against serious infections such as septicemia, respiratory tract infections, urinary tract infections, and intra-abdominal infections. However, due to its potential nephrotoxic and ototoxic side effects, careful monitoring of drug levels and renal function is essential during therapy. Its use is typically reserved for severe infections in which other antibiotics are ineffective (Narayanan and Lewis, 2023).

2.6. ANNOTATION PIPELINE FOR ANALYSIS OF WGS DATA

To create an annotation pipeline for our dataset, we must first identify the genes in the genome assemblies, that are related to antimicrobial resistant properties. Our aim is to focus on specific and important gene information so that the prediction of antimicrobial resistance can be made based on regions, sequences and variants that determine it. Firstly, we have to create a comprehensive database of antimicrobial resistance genes and two primary sources were utilized: the Comprehensive Antibiotic Resistance Database (CARD) and the Antimicrobial Resistance Gene Finder Plus (AMRFinderPlus). The CARD database was accessed and downloaded from [CARD](#), while the AMRFinderPlus database was obtained from [AMRFinderPlus](#).

2.6.1. COMPREHENSIVE ANTIBIOTIC RESISTANCE DATABASE (CARD)

The Comprehensive Antibiotic Resistance Database (CARD) is a robust bioinformatics resource that curates and annotates data on AMR genes. CARD integrates a vast array of molecular sequences associated with resistance mechanisms, providing detailed insights into how pathogens develop and propagate resistance to antibiotics. By systematically cataloging these genes and their variants into ontologies, CARD aids researchers in identifying resistance patterns and developing strategies to combat AMR. This database is crucial for advancing our understanding of resistance mechanisms and enhancing the accuracy of resistance gene detection in clinical and environmental samples (Alcock *et al.*, 2022).

To extract data from the CARD database, the Antibiotic Resistance Ontology (ARO) table was first retrieved. This table, published on October 2, 2023, is version 3.2.8. Moreover, this table contains all the genes along with their protein and DNA accession numbers and the associated mechanisms of resistance. It contains 5,178 resistance genes, whose proteins we downloaded using a bash script and using Protein accession numbers.

2.6.2. ANTIMICROBIAL RESISTANCE GENE FINDER PLUS (AMRFINDERPLUS)

AMRFinderPlus software and the accompanying database identify acquired antimicrobial resistance genes in bacteria, as well as known resistance-associated point mutations for several taxa. In AMRFinderPlus, the categorization of proteins into the ‘core’ and ‘plus’ groups serves to differentiate between distinct functionalities associated with antimicrobial resistance (Feldgarden *et al.*, 2021). Core proteins are those directly related to antimicrobial resistance mechanisms, such as beta-lactamases or efflux pumps, which are essential for conferring resistance to antibiotics. These proteins play a critical role in the pathogen’s ability to survive in the presence of antimicrobial agents. On the other hand, plus proteins encompass a broader spectrum of resistance factors that extend beyond traditional antimicrobial resistance. These include proteins involved in resistance to various environmental stresses and toxic substances such as biocides, heavy metals, heat, and acids, as well as those contributing to virulence. This categorization allows for a more comprehensive understanding of the genetic factors that contribute to the resilience and adaptability of microorganisms in diverse environments,

enhancing the ability to predict and manage antimicrobial resistance effectively (Feldgarden *et al.*, 2021).

To retrieve the reference gene catalog from the specified source, the AMRFinderPlus database was accessed through NCBI FTP site at [this link](#). The catalog, dated November 15, 2023, provides a comprehensive dataset in version 3.11, which includes detailed information on antimicrobial resistance genes. The data contains columns for gene names, associated RefSeq and Genbank accession numbers, scope of each gene (core or plus), specific resistance mechanisms, etc. Using a bash script, we downloaded 8.071 core proteins and 1.115 plus proteins, so we created an Antimicrobial peptides (AMPs) concatenated database of both CARD and AMRFinderPlus databases, resulting in 9.694 core and 1.084 plus proteins.

2.6.3. ANTIMICROBIAL PEPTIDES (AMPs) CLUSTERING

Organizing Antimicrobial peptides (AMPs) into clusters using the CD-HIT tool involves grouping peptide sequences that are similar to each other (Fu *et al.*, 2012). CD-HIT, a widely-used bioinformatics tool, efficiently reduces redundancy by clustering sequences that meet a specified similarity threshold. This process is crucial in our antimicrobial research, as it helps identify unique AMPs, analyze their diversity and help us with the identification of these AMPs families in genome assemblies.

To continue our analysis, we concatenated all the downloaded protein FASTA files into a single file and we ran CD-HIT tool, to identify protein families, based on homology. CD-HIT produced a file listing all IDs per cluster and a FASTA file of representative sequences. We provided as input the total protein sequences (both core and plus, totaling 10.778) and we used the arguments `-M 0` for memory usage, `-c` for sequence identity threshold and `-aL` for defining alignment coverage threshold. Afterward, we ran CD-HIT several times with different parameters to determine the best thresholds suitable for our protein database. The clustered results are shown in Table 2.2.

Based on these results and on the article of Zilhao *et al.*, 1988, the *tet* genes (*tetM*, *tetO*) can coexist in the genome, so if we put them in the same cluster we should keep them alone one of the two if both exist. Therefore, we choose the strict clustering of 0.9/0.9, resulting in 2.454 clusters, so that we achieve better results with Basic Local Alignment Search Tool (BLAST). A lower threshold would group these two genes into one representative.

<code>-c</code> Value	<code>-aL</code> Value	No of clusters
0.7	0.7	1.851
0.8	0.7	2.085
0.9	0.7	2.431
0.7	0.8	1.856
0.8	0.8	2.088
0.9	0.8	2.434
0.7	0.85	1.862
0.8	0.85	2.096
0.85	0.85	2.253
0.9	0.85	2.441
0.7	0.9	1.876
0.8	0.9	2.108
0.9	0.9	2.454

Table 2.2 | CD-HIT results using different clustering parameters. We performed multiple CD-HIT runs to find the best threshold for clustering Antimicrobial peptides.

2.6.4. AMPS DATABASE AND ALIGNMENT

In this step, we wrote a snakemake pipeline, which takes the most representative protein sequences derived from CD-HIT and make a BLAST compatible database (`makeblastdb` command). Moreover, we extended this pipeline to run BLASTx routine for each assembly, for finding matching sites with these AMPs proteins. The result is a file containing the name of the contig to which a match was found, the name of the aligned protein, the coordinates of the match on these two sequences, the percent identity, the e-value, the alignment length, the alignment of matched sequences, etc. In BLASTx results, alignment length is the total number of amino acid positions (including matches, mismatches, and gaps) involved in the alignment between the query sequence (translated assemblies) and the subject protein sequence.

Given the coordinate positions of BLAST, partially aligned regions and duplicate records must be filtered to avoid multiple matches of a protein in a genome. Based on the referenced criteria and additional tests, we calculated the global percent identity of each match ($(\text{alignment_length} - \text{gaps}) / \text{protein_length} * 100$). That's because the Percent identity index is calculated by dividing with the length of the query alignment, i.e. the assembly region. But this is a problem in our analyses, in case a protein is partially aligned or only in a part of it. For this, we introduced into our analysis the global percent identity index, which we divided by the total length of the protein (subject sequence). After, we retained all records where the protein matched the contig in at most its first 5 amino acids. We then filtered out records with a global percent identity lower than 70% and a percent identity lower than 70%. For any potential duplicate records remaining despite the filtering, we kept the one with the lowest *e-value*. This filtering process was implemented in a Python script.

The next step is to obtain the upstream sequences of the alignment region. To achieve this, we used the `Biostrings` R package to load assemblies and extract the 300 nucleotides upstream of the start of the alignment, i.e., from the 5'-end. For reverse strand alignments, we extracted the same sequence but transformed it using the reverse complement method. Finally, we saved these records in a FASTA file for each assembly, and in a second file, we recorded the assembly sequences of the subject alignment protein region (derived from BLASTx).

2.6.5. DETERMINATION OF NON-TRANSLATED GENES

The next step of the annotation is to introduce the ribosomal RNA information into our analysis. As mentioned in the Introduction, some antibiotics target ribosomal RNA in order to alter the process of bacterial protein synthesis and thus lead the cell to die. Therefore, extending the methodology to include 5S, 16S and 23S ribosomal copies of bacteria will provide important information in determining antimicrobial resistance phenotype based on genotypic data.

To match non-translated genes, like 5S, 16S and 23S rRNAs in genome assemblies, we extended Snakemake annotation pipeline to follow a specific set of steps. Firstly, we obtained the reference genomes from ESKAPEE pathogens and we downloaded complete sequences of 5S, 16S and 23S rRNA genes from the Nucleotide NCBI database (Table 2.3). Then, we made BLAST databases for these genes separately, one rRNA gene per database, and we used BLASTn routine for finding alignments between these genes and reference genomes.

Bacteria possess multiple copies of rRNA genes, so we anticipated that BLASTn would return several alignments. To manage this, we wrote a bash shell command to isolate the reference genome alignment sequence into a file and then split this file into multiple files, each containing a single sequence. We proceeded by creating a separate BLAST database for each new file for each genus and ran the BLASTn routine between each new database and each assembly. Continuing the analysis with a Python script, we loaded BLASTn results into a dictionary, where the keys were concatenations of the rRNA gene, genus, and copy number (e.g., `16s_Escherichia_coli_copy1`), and the values were data frames of the results. For each copy, we recorded

Bacterial Name	Strain	NCBI ID	rRNA Copy
Acinetobacter baumannii	strain ATCC 17978	NR_075445.1	5S
Enterobacter cloacae	strain ATCC 13047	NR_103226.1	5S
Enterococcus faecium	strain Aus0004	NR_103401.1	5S
Escherichia coli	strain Sakai O157:H7	NR_075284.1	5S
Klebsiella pneumoniae	strain ATCC 700721	NR_075468.1	5S
Pseudomonas aeruginosa	strain PAO1	NR_075116.1	5S
Staphylococcus aureus	strain N315	NR_075288.1	5S
Acinetobacter baumannii	strain K09-14	NZ_CP043953.1	16S
Enterobacter cloacae	strain ATCC 13047	NR_102794.2	16S
Enterococcus faecium	–	AY172570.1	16S
Escherichia coli	–	J01859.1	16S
Klebsiella pneumoniae	strain NB60 114	NZ_AZAP01000114.1:147	16S
Pseudomonas aeruginosa	strain DSM 50071	NR_026078.1	16S
Staphylococcus aureus	strain ATCC 12600	NR_118997.2	16S
Acinetobacter baumannii	strain ATCC 17978	NR_076472.1	23S
Enterobacter cloacae	strain ATCC 13047	NR_103067.1	23S
Enterococcus faecium	strain Aus0004	NR_103056.1	23S
Escherichia coli	strain K09-14	NR_075284.1	23S
Klebsiella pneumoniae	strain MGH 78578	NR_076493.1	23S
Pseudomonas aeruginosa	strain PAO1	NR_076166.1	23S
Staphylococcus aureus	strain N315	NR_076325.1	23S

Table 2.3 | Bacterial strains used to derive 5S, 16S and 23S rRNA copies. This table lists the bacterial strains from which the 5S, 16S, and 23S rRNA genes were derived. It includes the bacterial name, strain identification, NCBI accession number for the rRNA gene, and specifies which rRNA copy (5S, 16S, or 23S) is associated with each strain. The reference genomes from ESKAPEE pathogens and their rRNA sequences were obtained from the NCBI database to facilitate BLASTn alignment and analysis of rRNA genes in genome assemblies.

its name and the highest bit-score value in a data frame, sorting this data frame in descending order. Starting with the copy with the highest bit-score, we saved the record with the query ID and start & end coordinates in a separate data frame. For the copy with the second highest score, we removed records with the same coordinates and query ID as the previous one and kept the next best hit. This process was repeated for the remaining rRNA copies. For copies without sequences, we added the ‘X’ character, matching the length of the initial rRNA sequence, and concatenated all these sequences into a file as an entire sequence. Additionally, we filtered out records with overlapping start and end coordinates or coordinates within the range of previous hits.

In conclusion, the annotation pipeline gave us 3 different kinds of data for each assembly: protein data associated with antimicrobial resistance (protein data), nucleotide data from the upstream sequences of the alignment (upstream sequences), and rRNA sequence data of the contained copies (rRNA data). To have all these data in the same format, we should first add as many proteins that have not been found out of the total 2,454 in the protein data and as many segments that have not been found in the upstream sequences, so that we have for each genome the same number of records. Not found sequences will be represented by the character ‘X’, with the same length as the corresponding protein, or 300 if it is an upstream segment. Additionally, the order of these records should be the same, so with a Python script we sorted the files to have the same match across all assemblies, which will help feed this data into machine learning models, in the next section.

2.7. BUILDING AND EVALUATING MACHINE LEARNING MODELS

In this section, we will feed the annotated data (including the upstream nucleotide and protein sequences, as well as the rRNA copies) into machine learning algorithms, to get prediction of the phenotype (Susceptible, Resistant), based on these input sequences.

Regression models tend to be less stable when applied to datasets with a large, sparse input space and

correlated features, such as those involving drug resistance, which can be influenced by multiple antimicrobial resistance genes (Devika *et al.*, 2016). To enhance model performance, penalized regression techniques like LASSO and Ridge regression can be employed to reduce the input space and select the most important features. However, these methods assume that input features are uncorrelated. In datasets with correlated features, this assumption leads to distorted feature selection and, consequently, less reliable model interpretation.

In contrast, decision trees inherently handle correlated features more effectively, as correlation does not affect the feature selection process. Random forest models, which construct hundreds to thousands of decision trees using different subsets of the input space, offer a more robust identification of important features. XGBoost, an advanced gradient boosting algorithm for decision trees, is adept at handling correlated inputs and supports sparse datasets, making it well-suited for our data set. This algorithm can extract critical features, providing valuable insights into the decision-making process. While XGBoost supports multi-class classification, we developed separate, independent binomial models (resistant vs susceptible) for the antibiotic of interest.

As stated above, the input for each model was the annotated data derived from the previous section. This data includes each sequence separately, so we have 2,454 Antimicrobial peptides sequences, 2,454 upstream sequences and 3 rRNA sequences. The output was binomial with label resistant (1) or susceptible (0) to the Gentamicin antibiotic, because we have excluded Not detected and Intermediate categories. Also, we applied several encoding methods in the total input vector, as described below and in our data set we have 10,376 susceptible and 5,325 resistant isolates (Figure 2.2). We then splitted the data set to 80% training and 20% testing, before the algorithm's training.

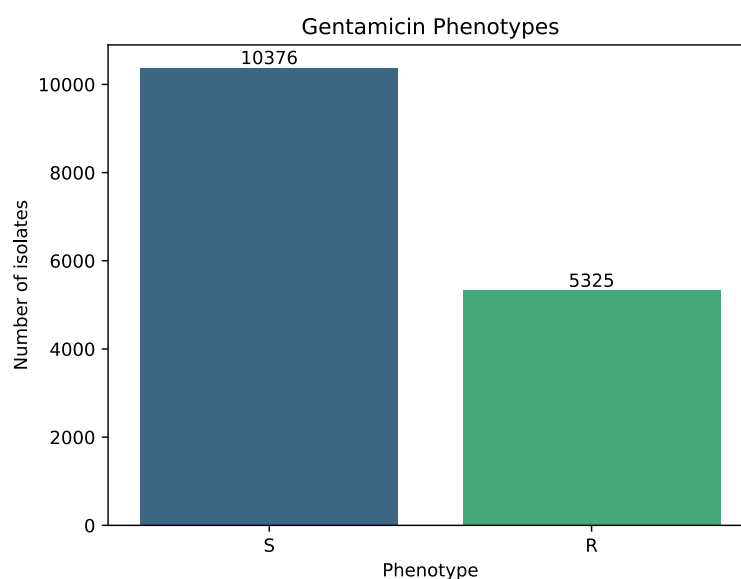


Figure 2.2 | Distribution of Gentamicin Phenotypes. This figure displays the number of susceptible (*S*) or resistant (*R*) isolates to the Gentamicin antibiotic. The dataset comprises 10,376 susceptible isolates and 5,325 resistant isolates.

2.7.1. ENCODING METHODS

2.7.1.1. *k*-MER ENCODING

k-mer algorithm is pivotal in the analysis of Antimicrobial Resistance (AMR) sequences due to their efficiency in identifying and characterizing resistance genes in large genomic datasets. A *k*-mer refers to a substring of length *k* within a DNA or protein sequence. By breaking down sequences into *k*-mers and counting them, algorithms

can efficiently compare these short segments across different genomes to identify patterns and anomalies indicative of antimicrobial resistance. K -mer encoding facilitates rapid scanning and classification of genomic sequences, enabling the identification of known resistance genes and the detection of novel resistance-associated mutations. For instance, tools like KmerResistance, matches k -mers from sequenced isolates against a database of resistance genes to predict phenotypic resistance profiles (Clausen *et al.*, 2018), indicating the power of sequence segmentation in bioinformatics using k -mer algorithm.

k -mer encoding is implemented using a character-level **CountVectorizer** from the scikit-learn library (Pedregosa *et al.*, 2011), which transforms the sequences into numerical representations based on k -mer frequency. The analysis proceeds by fitting the vectorizer to the sequence data, transforming it into sparse matrices, and then converting these matrices into dense arrays. The data is filtered to exclude sequences with minimal information, reshaped to its original structure, and concatenated to avoid padding issues. Finally, the processed k -mer data is saved as a NumPy array (Harris *et al.*, 2020) for further analysis. This methodology ensures efficient and accurate encoding of genomic sequences, facilitating downstream analysis and machine learning applications.

2.7.1.2. ONE-HOT ENCODING

One-hot encoding is a widely used technique in data preprocessing for converting categorical variables into a binary vector representation that machine learning algorithms can efficiently utilize. Each category value is converted into a unique binary vector, where only one element is ‘hot’ (set to 1) and all other elements are ‘cold’ (set to 0). This encoding scheme is particularly useful for categorical variables where there is no ordinal relationship between the categories, as it ensures that the model treats each category as distinct and equidistant from the others.

However, one-hot encoding can lead to high-dimensional data, especially when dealing with features that have a large number of categories. This increase in dimensionality can make the model more complex and computationally expensive to train. Despite this, one-hot encoding remains a fundamental technique in machine learning preprocessing, as it effectively preserves the categorical nature of the data without introducing spurious relationships.

One-hot encoding is implemented using the **OneHotEncoder** function from the scikit-learn library, to encode categorical features as a one-hot numeric array. After encoding each sequence, the resulting binary vectors are flattened and concatenated to create a single continuous vector. This is crucial for providing the data in a format suitable for machine learning models.

2.7.1.3. PSSM ENCODING

The Position - Specific Scoring Matrix (PSSM) encoding methodology was employed to enhance the prediction of antimicrobial resistance by leveraging evolutionary information from protein sequences. This approach involved several key steps to ensure that the resistance profiles of proteins were accurately captured and utilized in the training process. First, we collected resistance-related proteins from each assembly and performed multiple sequence alignments, one per same protein, using Clustal Omega (Sievers and Higgins, 2018). These alignments helped in identifying conserved regions and variations across the protein sequences. From these multiple alignments, we constructed PSSM matrices for each protein. The PSSM matrices capture the probability of each amino acid occurring at each position in the alignment, providing a comprehensive representation of the evolutionary information embedded in the sequences. Pseudocounts are added to avoid zero frequencies and ensure a more robust estimation and the adjusted frequencies are computed accordingly. Also, background frequencies for each amino acid are predefined based on typical amino acid distributions.

During the testing phase, the protein sequences were passed through their corresponding multiple alignments, in order to be aligned and have the corresponding length and gaps, and subsequently through the PSSMs. This process allowed us to transform the protein sequences into 1D score matrices. Each score in the matrix represented the likelihood of a specific amino acid occurring at a particular position in the sequence, based on the PSSM. This transformation enabled the model to utilize the encoded evolutionary information effectively. By integrating PSSM encoding, we aimed to capture the subtle variations and conserved patterns in protein sequences that are indicative of antimicrobial resistance. This methodology ensured that the model was trained on enriched data, leading to accurate predictions and a better understanding of the resistance mechanisms at the molecular level.

2.7.2. XGBOOST MODEL TRAINING AND TESTING

The XGBoost algorithm, short for Extreme Gradient Boosting, is a powerful and highly efficient implementation of gradient boosting designed for speed and performance. Developed by Chen and Guestrin, 2016, XGBoost has become a popular choice in machine learning competitions and real-world applications due to its ability to handle a wide variety of predictive modeling tasks with high accuracy. It works by building an ensemble of decision trees sequentially, where each new tree attempts to correct the errors made by the previous trees. This iterative process optimizes a loss function and improves the overall predictive power of the model. XGBoost incorporates several advanced features, such as regularization to prevent overfitting, parallel processing to speed up computation, and handling of missing data. Its capability to handle large-scale datasets and its robustness to different types of data make it particularly well-suited for tasks like classification and regression, including challenging problems like antimicrobial resistance prediction.

The objective function in XGBoost consists of two main components: a Loss function that measures the predictive accuracy of the model, and a Regularization term that controls the complexity of the model to prevent overfitting. The combination of these components in the objective function helps XGBoost to not only achieve high accuracy by fitting the data well but also maintain generalizability by avoiding overfitting through regularization. We trained the XGBoost algorithm using various binary and multiclass objective functions, specifically: `binary:logistic`, `binary:logitraw`, `binary:hinge`, `multi:softmax` and `multi:softprob` with a focus on two classes (S and R). The `binary:hinge` and `multi:softmax` functions return discrete predictions of 0 or 1, while the other functions return the predicted probabilities for each class (0 and 1).

2.7.3. RANDOM FOREST TRAINING AND TESTING

The random forest algorithm is a popular ensemble machine learning technique used for classification and regression tasks (Breiman, 2001). It operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. This method is known for its robustness and ability to handle large datasets with higher accuracy. The random forest algorithm works by randomly selecting subsets of features and data samples, ensuring that each tree in the forest is trained on a different portion of the data, which helps in reducing overfitting and improving generalization (Breiman, 2001). In addition, it provides a measure of feature importance, making it a valuable tool for identifying the most significant variables in a dataset. Its ensemble approach of combining multiple decision trees helps in capturing the complexities and interactions in the data, leading to more reliable and accurate predictions. Scikit-learn (Pedregosa *et al.*, 2011) provides an easy-to-use implementation of the Random Forest algorithm through the `RandomForestClassifier` class.

2.7.4. HYPERPARAMETERS TUNING

Hyperparameter tuning involves optimizing the parameters that control the learning process and model behavior to achieve the best performance on a given data set. This process is crucial because the choice of hyperparameters can significantly impact the accuracy, overfitting, and generalization ability of the model.

Key hyperparameters in XGBoost include the learning rate, which determines the step size at each iteration, the number of trees or estimators, which affects the model's complexity, max depth, controlling the depth of each tree, and subsample, which dictates the fraction of data to be used for training each tree. Advanced methods like Bayesian optimization can also be employed to efficiently navigate the hyperparameter space. Proper tuning ensures that the XGBoost model is both accurate and robust, maximizing its predictive power.

First, we have applied the k -mer algorithm to the data for k equal to 3, 4 and 5 and saved the files for training the algorithms. Then, for each different k , we made hyperparameter tuning on the XGBoost algorithm, with the following hyperparameters:

$$\begin{aligned} \text{eta: } & [0.01, 0.1], \\ \text{max_depth: } & [5, 7, 9, 11, 13, 15, 17, 19, 21, 23], \\ \text{num_boost_round: } & [100, 200], \end{aligned}$$

All other parameters were kept default, evaluation was done during training, based on the test data set and early stopping was done when no improvement was seen in 10 consecutive XGBoost rounds.

In addition, the key hyperparameters for the Random Forest algorithm include the number of trees in the forest (n_estimators), the maximum depth of the trees (max_depth) and the number of features to consider when looking for the best split (max_features). These hyperparameters were tuned to identify the best combination that minimizes prediction error and maximizes the robustness of the model. So, for Random Forest, we conducted hyperparameter tuning with the following ranges:

$$\begin{aligned} \text{n_estimators: } & [100, 200, 400], \\ \text{max_depth: } & [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], \\ \text{bootstrap: } & [\text{True}, \text{False}], \end{aligned}$$

Other parameters were kept at their default values and the optimal hyperparameters were chosen based on the model's best accuracy performance. This tuning process ensures that the Random Forest model is well-calibrated to the specific characteristics of the dataset, thereby enhancing its predictive accuracy and generalization capability.

2.7.5. PERFORMANCE METRICS IN ANTIMICROBIAL RESISTANCE PREDICTION

When evaluating the performance of an ML model in predicting antimicrobial resistance, several key metrics are used. Some of them are: precision, recall, F1 score, macro average, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Understanding these metrics is crucial to assess how well the model distinguishes between susceptible (class 0) and resistant bacterial strains (class 1).

Accuracy. Accuracy is a fundamental metric used to evaluate the performance of a classification model. It measures the proportion of correctly classified instances among the total number of instances in a dataset. In simpler terms, accuracy indicates how often the model makes the right prediction:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

Precision. Precision is the ratio of correctly predicted positive observations to the total predicted positives.

It is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (2.2)$$

In the context of antimicrobial resistance prediction, precision indicates the proportion of strains predicted to be resistant that are actually resistant. High precision means that when the model predicts resistance, it is usually correct.

Recall. Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to the all observations in actual class 1. It is defined as:

$$Recall = \frac{TP}{TP + FN} \quad (2.3)$$

For our prediction, recall measures the proportion of resistant strains that are correctly identified by the model. High recall is crucial in this context because it ensures that most resistant strains are detected.

F1 Score. The F1 score is the harmonic mean of precision and recall and is used to balance the two metrics. It is defined as:

$$F1_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.4)$$

A high F1 score indicates a good balance between precision and recall, which is important for a reliable prediction model.

Macro Average. The macro average (macro avg) is the average of precision, recall, and F1 scores across all classes. It treats all classes equally, providing a single performance measure. In this problem, macro average is useful to get an overall sense of the model's performance across both susceptible and resistant classes.

True Positives, True Negatives, False Positives, and False Negatives:

- True Positives (TP): The number of resistant strains correctly identified as resistant.
- True Negatives (TN): The number of susceptible strains correctly identified as susceptible.
- False Positives (FP): The number of susceptible strains incorrectly identified as resistant.
- False Negatives (FN): The number of resistant strains incorrectly identified as susceptible.

These counts are fundamental to calculating precision, recall, and other metrics.

In the clinical setting, especially in antimicrobial resistance prediction, recall (sensitivity) is more critical than accuracy. The primary goal is to ensure that resistant strains are not missed, as failing to identify a resistant infection could lead to inappropriate treatment, prolonged infection, and the spread of resistance. False negatives (resistant strains classified as susceptible) can have severe consequences, including treatment failure and increased risk of complications. Although, accuracy is important in minimizing the number of false alarms (susceptible strains incorrectly classified as resistant), the clinical risk of a false negative outweighs the inconvenience of a false positive. Thus, improving recall ensures that most, if not all, resistant strains are detected, allowing appropriate and effective treatment regimens to be administered promptly. This approach helps prevent the spread of resistance and improves patient outcomes. In summary, while accuracy, recall, and other performance metrics all contribute to a comprehensive understanding of model performance, recall holds paramount importance due to the high stakes associated with undetected resistant infections.

2.7.6. k -FOLD CROSS – VALIDATION

Cross-validation is a statistical technique used to evaluate the performance and generalizability of a machine learning model. It involves partitioning the dataset into multiple subsets, training the model on some subsets while validating it on the remaining ones. This process helps to estimate how well the model will perform on an independent dataset, thus mitigating issues like overfitting and ensuring the robustness of the model. Cross-validation is particularly valuable when the amount of data is limited, as it maximizes the use of available data for both training and validation purposes.

One of the most common forms of cross-validation is k -fold cross-validation. In k -fold cross-validation, the data is divided into k equally sized folds or subsets. The model is trained k times, each time using a different fold as the validation set and the remaining $k-1$ folds as the training set. The performance metrics from each of the k runs are then averaged to produce a single estimation. This method ensures that every data point is used for both training and validation, which provides a more reliable estimate of the model's performance compared to a single train-test split. For instance, a study by Kohavi, 1995 demonstrated the effectiveness of k -fold cross-validation in evaluating the accuracy of predictive models across various domains.

Another widely used technique is stratified k -fold cross-validation. This method is particularly useful for imbalanced datasets, where certain classes are underrepresented. In stratified k -fold cross-validation, the folds are created such that each fold maintains the same proportion of classes as the original dataset. This ensures that the model is exposed to all classes during training and validation phases, thereby providing a more balanced evaluation. According to a study by Japkowicz and Stephen, 2002, stratified k -fold cross-validation significantly improves the evaluation of classifiers on imbalanced datasets.

3 | RESULTS

In this study, we utilized Machine learning to predict Antimicrobial Resistance from Whole Genome Sequencing data obtained from public databases. The focus was on seven common prevalent nosocomial pathogens, collectively known as ESKAPEE pathogens: *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter species* and *Escherichia coli*. These organisms were chosen due to their clinical significance and extensive representation in AMR-annotated samples.

3.1. DATA COLLECTION AND PRE-PROCESSING

Data were sourced from the Isolates Browser on the NCBI Pathogen Detection website (NDARO database), filtered for ESKAPEE pathogens and AST phenotypes. The data set included 13,827 isolates corresponding to 12,865 unique assemblies. BioSample IDs were used to retrieve antibiogram tables, and a comprehensive data frame was created by consolidating all antibiograms. We excluded assemblies with measurement units in millimeters and those with unbalanced read data, resulting in 13,159 unique assemblies. Furthermore, data from the NARMS website were processed, yielding 477 unique assemblies after filtering. Similarly, from the BV-BRC database, 103,012 records were initially obtained, filtered down to 99,944 records, resulting in 6,249 unique assemblies. Combining datasets from NDARO, NARMS, and BV-BRC, we achieved a final set of 19,132 unique assemblies after removing duplicates.

Quality control was conducted using the N50 metric, which indicates the contiguity and completeness of genome assemblies. Assemblies with low N50 values were flagged or removed from the analysis. N50 values were obtained from NDARO and calculated for NARMS and BV-BRC using the `Biostrings::N50` function. Exploratory analysis revealed that *Enterobacter* had the highest mean N50 value, while *Escherichia coli* had the lowest. After applying N50 cutoff values (50,000 general cutoff value and 20,000 for *Enterobacter* and *Staphylococcus*), 587 low-quality assemblies were removed, leaving 18,545 unique strains for further analysis. This meticulous data collection and quality control process ensured the reliability of the genomic data used for predicting antimicrobial resistance.

3.2. k -FOLD CROSS – VALIDATION

In the course of our study, we employed cross-validation techniques to evaluate the performance and reliability of our machine learning models. Cross-validation is a crucial method in machine learning to assess how well a model generalizes to an independent dataset. By partitioning the data into training and testing sets multiple times, cross-validation ensures that the model is tested on different subsets, providing a comprehensive evaluation.

One of the primary methods we used is stratified k -fold cross-validation and in our experiments, we divided the data set into $k = 3$ folds, ensuring that each fold contained an approximately equal proportion of resistant and susceptible cases. This approach not only provided a robust evaluation framework, but also helped mitigate potential biases due to class imbalance. The XGBoost model, which we fine-tuned using hyperparameter optimization, demonstrated consistent performance across all folds. These hyperparameters was the following:

$$\begin{aligned} \text{eta: } & [0.1], \\ \text{max_depth: } & [5, 7, 9, 11, 13, 15, 17, 19, 21, 23], \\ \text{num_boost_round: } & [200], \end{aligned}$$

The number of boosting rounds was set to 200 as an upper limit, with early stopping applied if there were no improvements in the validation metric for 10 consecutive rounds. The learning rate (eta) was fixed at 0.1, as the previous hyperparameter tuning indicated that this value yielded slightly better performance metrics. The model demonstrated an average accuracy of more than 97.6% across the folds, indicating a high level of predictive power. In addition, precision, recall, and F1 scores were calculated for each fold to provide a complete understanding of model performance.

We executed the model with k -mer values of 3, 4, and 5, performing hyperparameter tuning for each instance using the specified hyperparameters. Through hyperparameter tuning with stratified k -fold cross-validation, we determined that the optimal max depth value was 15 for a k -mer value of 4. Training the model with these hyperparameters resulted in an accuracy of **98.29%**. The confusion matrix and a comprehensive classification report of the cross-validation process, detailing all relevant metrics, are presented in Figure 3.2 and in Table 3.1 are shown the best scores and hyperparameters for each instance. Figures 3.1 and 3.3 illustrate the confusion matrix and the classification report for k -mer values of 3 and 5, respectively.

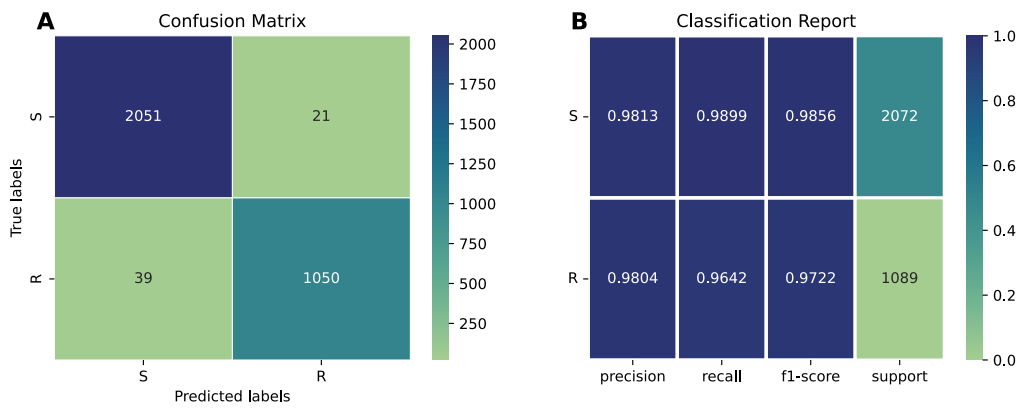


Figure 3.1 | Gentamicin predicted Phenotypes for k -mer size 3, using XGBoost model and Cross-Validation. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class.

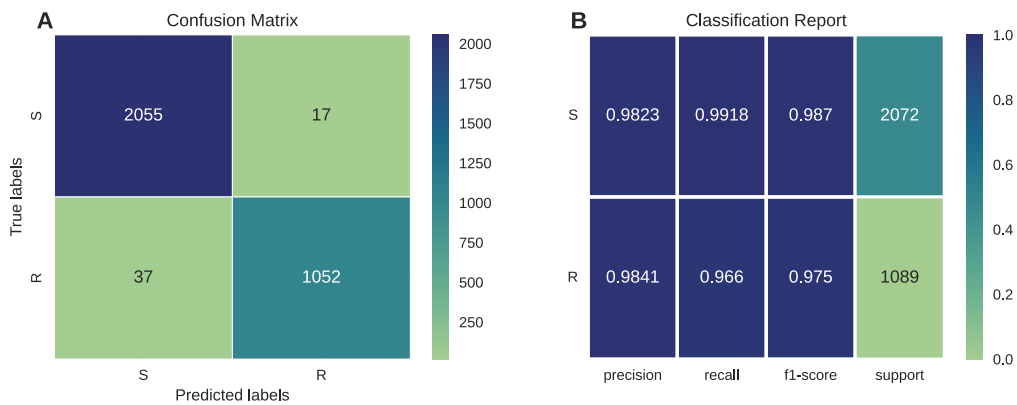


Figure 3.2 | Gentamicin predicted Phenotypes for k -mer size 4, using XGBoost model and Cross-Validation. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class.

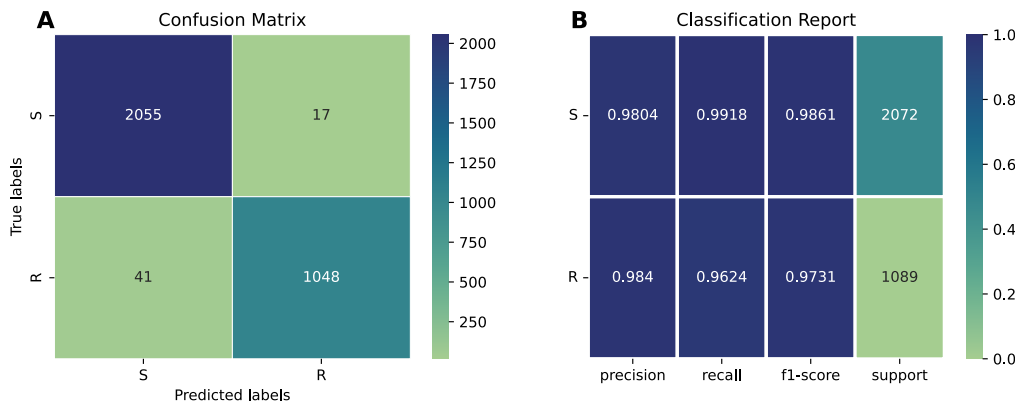


Figure 3.3 | Gentamicin predicted Phenotypes for k -mer size 5, using XGBoost model and Cross-Validation. (A) Confusion matrix of the model’s predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class.

Stratified k -fold Cross Validation (XGBoost)					
k -mer value	Hyperparameters	Accuracy	Precision	Recall	F1 Score
3	Max depth: 11				
	Eta: 0,1	0.9810	0.9804	0.9642	0.9722
	Boost rounds: 62				
4	Max depth: 15				
	Eta: 0,1	0.9829	0.9840	0.9660	0.9750
	Boost rounds: 59				
5	Max depth: 21				
	Eta: 0,1	0.9816	0.9840	0.9623	0.9731
	Boost rounds: 59				

Table 3.1 | k -mer data set used with stratified k -fold cross-validation on the XGBoost model. It is showing the performance metrics of accuracy, precision, recall, and F1 score across different k -mer values and hyperparameters.

3.3. XGBOOST PERFORMANCE METRICS

In this section, we provide a detailed analysis of the XGBoost model’s performance in predicting Antimicrobial Resistance (AMR) using different hyperparameters and k -mer values. We will also provide key metrics such as accuracy, precision, recall and the F1 score. The hyperparameters tuned during this process include the learning rate (eta), the maximum depth of the trees, and the number of boosting rounds. Specifically, the hyperparameters tested were:

eta : [0.01, 0.1],
 max_depth : [5, 7, 9, 11, 13, 15, 17, 19, 21, 23],
 num_boost_round : [100, 200],
 $objective$: ‘multi:softmax’

k -mer Value 3

For k -mer value 3, we conducted hyperparameter tuning (without cross-validation) to determine the optimal configuration for the XGBoost model. The results indicated that the best performance was achieved with a learning rate of 0.1, a maximum depth of 21 and 100 boosting rounds. This configuration yielded an accuracy of 98.29%. The precision, recall and F1 score were consistently high, demonstrating the model’s robustness and reliability. Detailed metrics, including the confusion matrix and classification report for k -mer value 3, are presented in Figure 3.4.

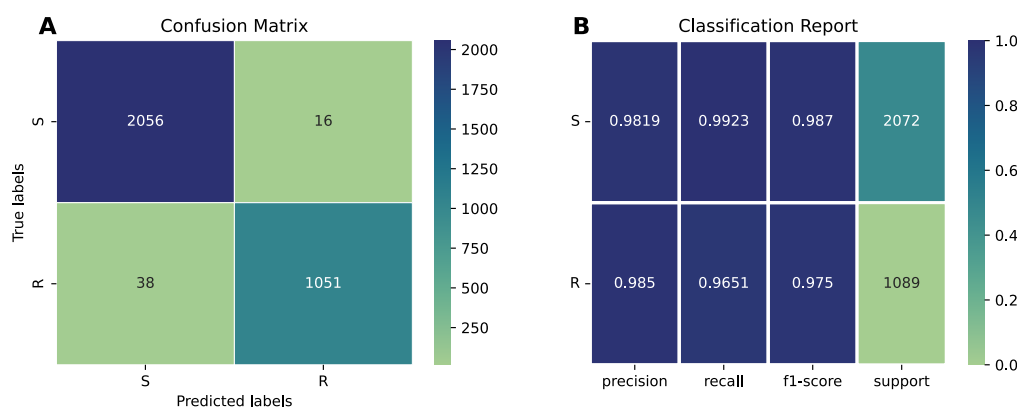


Figure 3.4 | Gentamicin predicted Phenotypes for k -mer size 3, using the XGBoost model. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class.

k -mer Value 4

For k -mer value 4, the hyperparameter tuning process identified the optimal maximum depth value as 19, with a learning rate of 0.1 and 100 boost rounds. Training the model with these hyperparameters resulted in an accuracy of 98.35%, showcasing the highest performance among the tested k -mer values. The confusion matrix and the comprehensive classification report, which details all relevant metrics, are shown in Figure 3.5. This high level of accuracy, combined with strong precision, recall and F1 scores, highlights the model's effectiveness in predicting AMR with this specific k -mer value.

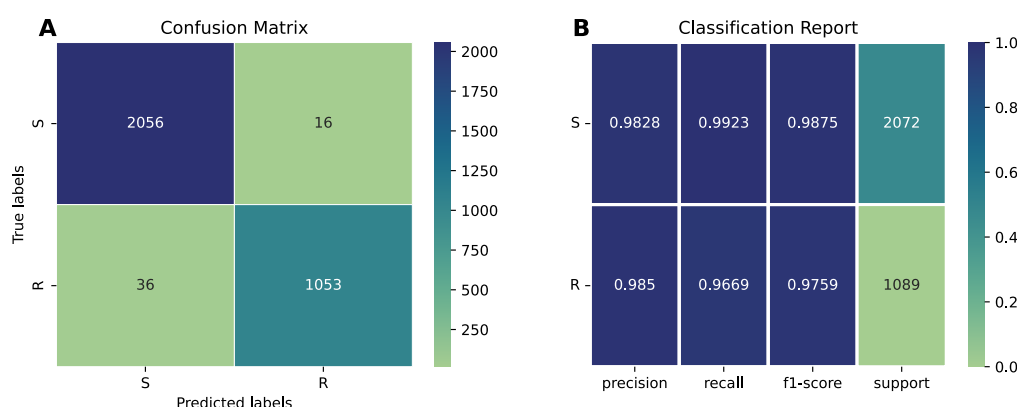


Figure 3.5 | Gentamicin predicted Phenotypes for k -mer size 4, using the XGBoost model. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class.

k -mer Value 5

For k -mer value 5, the optimal hyperparameters were determined to be a learning rate of 0.1, a max depth of 19, and 100 boosting rounds. The model achieved an accuracy of 98.196% under these conditions. While slightly lower than the performance with k -mer value 4, the results still indicate a high predictive power. The precision, recall and F1 scores were also evaluated to ensure a comprehensive understanding of the model's performance. The confusion matrix and classification report for k -mer value 5 are depicted in Figure 3.6.

We conducted another round of hyperparameter tuning on the dataset with a k -mer value of 5, altering the objective function from `multi:softmax` to `multi:softprob`, `binary:logistic` and `binary:hinge`. This iteration revealed that each objective function yielded different optimal hyperparameters, demonstrating the necessity of tailoring hyperparameter selection to the specific objective function to achieve the best performance. The best performance was achieved with an eta (learning rate) of 0.1, a max depth of 19, and 55 boosting

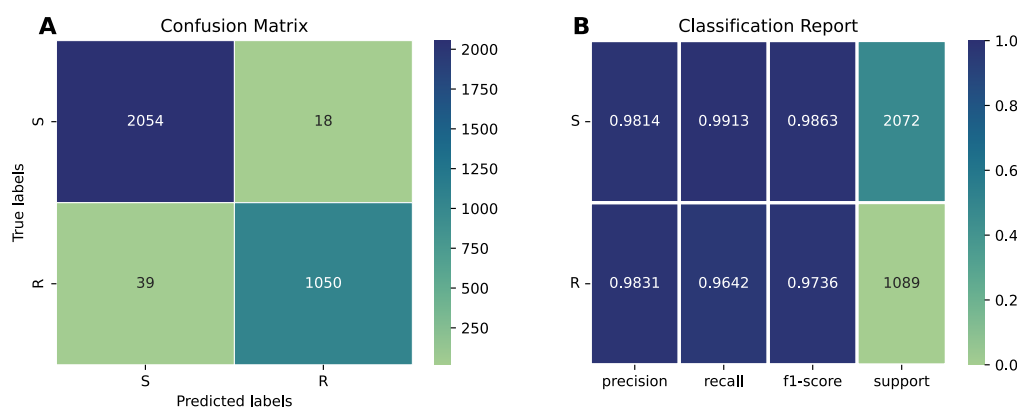


Figure 3.6 | Gentamicin predicted Phenotypes for k -mer size 5, using the XGBoost model. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class.

rounds, using both the `multi:softmax` and `multi:softprob` objectives, resulting in an accuracy of 98.196%. These objectives are designed for multi-class classification, with `multi:softmax` outputting the predicted class directly, while `multi:softprob` provides probabilities for each class. Additionally, using the `binary:logistic` objective, which is tailored for binary classification, the model achieved an accuracy of 98.13% with a maximum depth of 17 and 64 boosting rounds. Finally, the `binary:hinge` objective, which outputs binary predictions, achieved an accuracy of 98.07% with a maximum depth of 9 and 69 boosting rounds. These results underscore the robustness of the model and high predictive accuracy across different objectives and hyperparameter settings.

In Figure 3.7A, the Recall and Accuracy Scores are plotted against the probabilities thresholds. The recall curve (in green) demonstrates the model's ability to identify all true positive cases (resistant strains) across varying thresholds. The accuracy curve (in blue) reflects the overall correctness of the model's predictions. The intersection of these curves suggests an optimal threshold where the balance between recall and accuracy is maximized, indicating high performance in distinguishing between susceptible and resistant instances. Figure 3.7B provides a comprehensive of the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold settings. The area under the ROC curve (AUC) is 0.99, highlighting the excellent discriminative ability of the model. A high AUC value close to 1.0 indicates that the model effectively differentiates between the classes, with a minimal rate of false positives, confirming the robustness of the model in predicting antimicrobial susceptibility and resistance.

Figure 3.7C illustrates the predicted probabilities of antimicrobial susceptibility (Class 0) and resistance (Class 1) using the `multi:softprob` objective function in the XGBoost model. The density plot shows two distinct peaks, indicating the confidence of the model in its predictions. The left peak, shown in orange, represents the probability distribution for Class 0 (susceptible), with most probabilities clustering near 0. The right peak, shown in blue, represents the probability distribution for Class 1 (resistant), with most probabilities clustering near 1. This clear separation between the two classes suggests that the model performs well in distinguishing between resistant and susceptible instances, providing high predictive accuracy and no ambiguity. The predicted probabilities for the true class are significantly higher than those for the other classes and as a result, the class with the highest probability is always the correct one.

As we mentioned above, eliminating False Negatives, where resistant strains are incorrectly predicted as susceptible, is crucial in clinical practice to prevent ineffective treatments. Administering antibiotics to which the bacteria are resistant can lead to treatment failures, prolonged infections, and increased morbidity and mortality (Centers for Disease Control & Prevention, 2019). This not only jeopardizes patient outcomes, but also contributes to the spread of resistant strains, exacerbating the issue of antimicrobial resistance. In a

clinical setting, the correct identification of resistant strains ensures that patients receive the most effective treatment promptly, reducing the risk of complications and the likelihood of resistant bacteria proliferating and transmitting to others (Centers for Disease Control & Prevention, 2019).

To eliminate False Negatives (where resistant strains are misclassified as susceptible), the threshold for predicting a strain as resistant can be lowered. For example, instead of using the default 0.5 probability threshold, the threshold for classifying a strain as resistant can be reduced to 0.3. This means that any strain with a resistance probability greater than 0.3 would be classified as resistant. This approach increases the sensitivity of the model, ensuring that fewer resistant strains are incorrectly classified as susceptible. Adjusting these thresholds involves evaluating the recall and accuracy trade-offs. By prioritizing recall (the ability to correctly identify resistant strains), the model minimizes the risk of False Negatives, ensuring that more resistant strains are detected and appropriately treated. This adjustment can be fine-tuned based on the clinical context and the acceptable balance between recall and accuracy to optimize treatment outcomes and mitigate the spread of antimicrobial resistance.

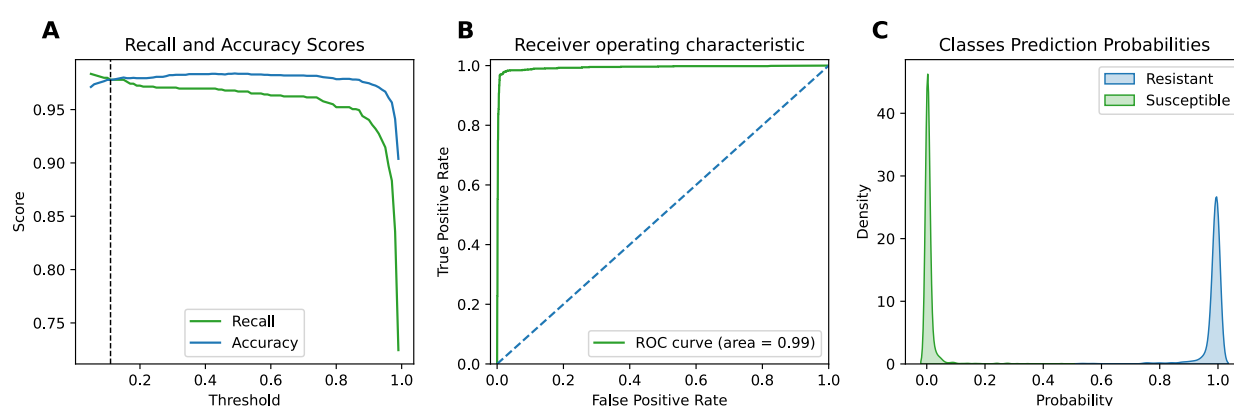


Figure 3.7 | Performance Metrics for XGBoost Model in Predicting Antimicrobial Resistance using softprob objective function. (A) Recall and Accuracy Scores plotted against the probabilities thresholds, demonstrating the balance between the model's ability to identify true positives and its overall prediction correctness. (B) Receiver Operating Characteristic (ROC) curve with an Area Under the Curve (AUC) of 0.99, indicating the model's excellent discriminative power between susceptible and resistant instances. (C) The density plot illustrates the predicted probability distributions for susceptible (Class 0) and resistant (Class 1) categories using the XGBoost model. The plot shows two distinct peaks, with susceptible probabilities clustering near 0 and resistant probabilities clustering near 1. This clear separation indicates the model's high confidence and accuracy in distinguishing between resistant and susceptible instances.

On the other hand, the results from the XGBoost model using one-hot encoding demonstrate high performance across various evaluation metrics. With best hyperparameters set to a maximum depth of 5, number of boost rounds of 200, and a learning rate of 0.1, the model achieved an impressive accuracy of **98.03%**. Breaking down the results further, the precision, recall, and F1-score for both classes are consistently high (Figure 3.8). For the susceptible class, the precision and recall are both at 0.982 and 0.988, respectively, leading to an F1-score of 0.985. For the resistant class, the precision and recall are 0.976 and 0.966, respectively, with an F1-score of 0.971. These metrics indicate that the model is highly effective in distinguishing between susceptible and resistant strains with minimal misclassification. Additionally, the macro and weighted averages for precision, recall, and F1-score are all around 0.98, further confirming the model's robustness and reliability.

Finally, the results from the XGBoost model using PSSM encoding demonstrate slightly lower performance metrics than those observed with *k*-mer and one-hot encoding. With the best hyperparameters set to a maximum depth of 19 and number of boost rounds of 100, the model achieved an accuracy of 97.63%. Delving into the specifics, the precision, recall, and F1-score for both classes are quite high. For the susceptible class, the precision and recall are 0.977 and 0.986, respectively, resulting in an F1-score of 0.982. For the resistant class, the precision is 0.973 and the recall is 0.956, yielding an F1-score of 0.965. Additionally, the macro and weighted averages for precision, recall, and F1-score are all between 0.97 and 0.98, indicating consistent

performance across classes. This high level of performance suggests that the XGBoost model, with the specified hyperparameters and PSSM encoding, remains a robust tool for predicting antimicrobial susceptibility and resistance, providing valuable insights for clinical practice and the management of antimicrobial resistance.

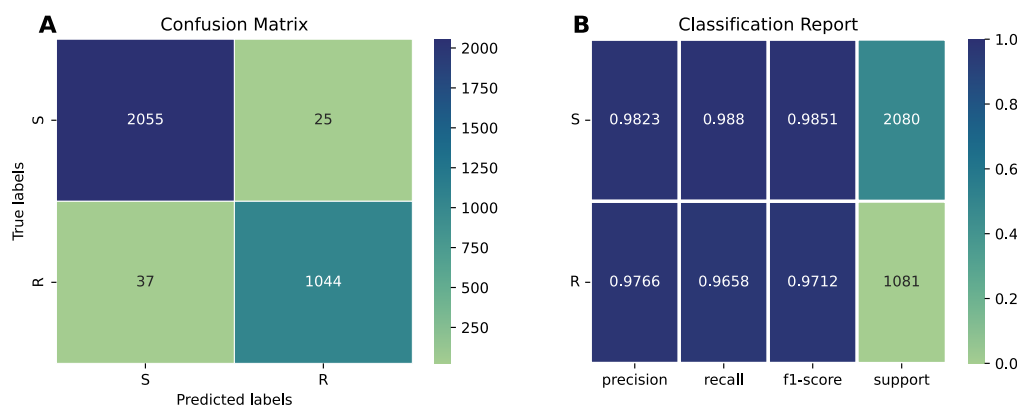


Figure 3.8 | Gentamicin predicted Phenotypes using XGBoost algorithm and one-hot encoding. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 98.03%.

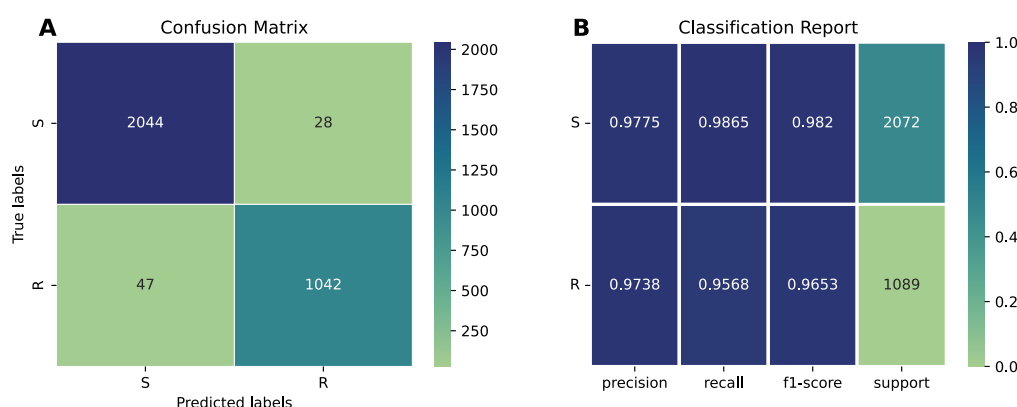


Figure 3.9 | Gentamicin predicted Phenotypes using XGBoost algorithm and PSSM encoding. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 97.63%.

3.4. RANDOM FOREST PERFORMANCE METRICS

This section provides the performance of the Random Forest model in predicting Antimicrobial Resistance (AMR) using different hyperparameters and different k -mer values, through key metrics such as accuracy, precision, recall and the F1 score. The hyperparameters tuned during this process include the number of estimators, the maximum depth of the trees, and a boolean value for enabling bootstrapping statistical technique. Specifically, the hyperparameters tested were the following, while all the others remained default:

$$\begin{aligned}
 n_estimators: & [100, 200, 400], \\
 max_depth: & [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], \\
 bootstrap: & [True, False]
 \end{aligned}$$

k -mer Value 3

For k -mer value 3, the results of the fine tuning indicated that the best performance was achieved with a number of estimators of 100, a maximum depth of 40 and bootstrapping. This configuration yielded an accuracy of **98.48%**, outperforming the prediction accuracy of the corresponding XGBoost model for k -mer size equal to 3. Detailed metrics, including the confusion matrix and classification report, are presented in Figure 3.10.

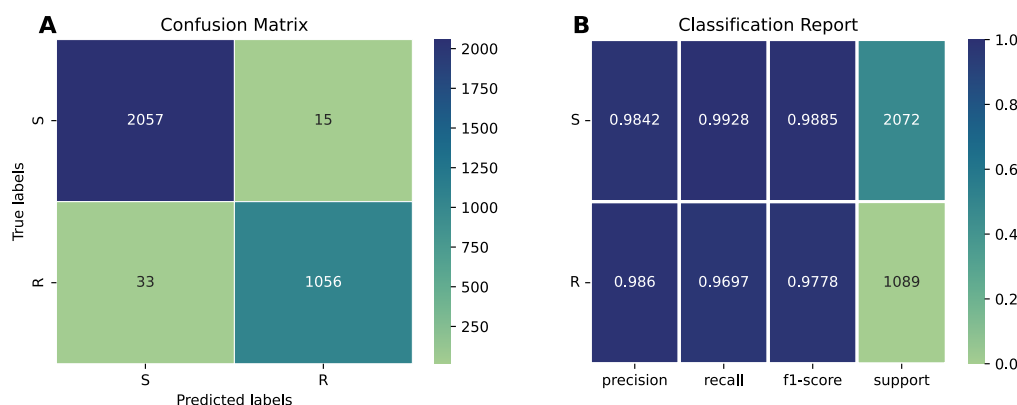


Figure 3.10 | Gentamicin predicted Phenotypes for k -mer size 3, using Random Forest algorithm. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 98.48%.

k -mer Value 4

For k -mer value 4, the results of the fine tuning indicated the same accuracy with k -mer size 3 and the used hyperparameters was: number of estimators of 200, a maximum depth of 50 and bootstrapping. Comparing these two cases for k -mer sizes 3 and 4, it turns out that it is preferable to use the model trained on k -mer size 3 data, due to less complexity and used parameters. Detailed metrics, including the confusion matrix and classification report, are presented in Figure 3.11.

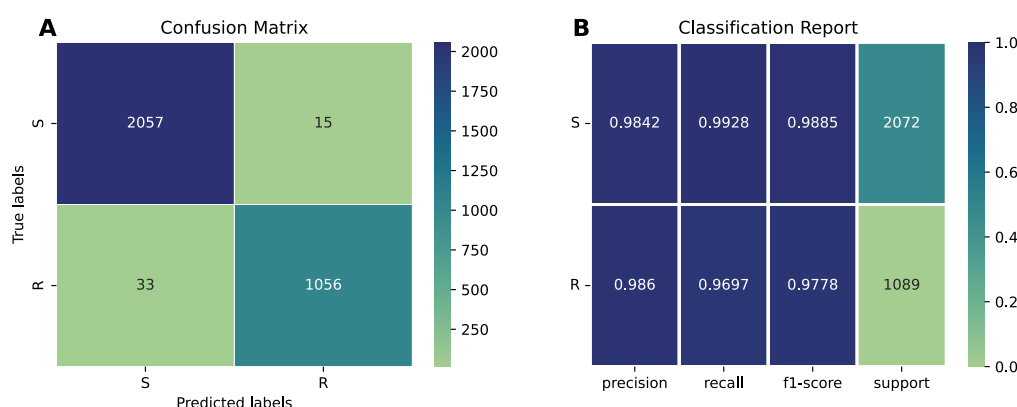


Figure 3.11 | Gentamicin predicted Phenotypes for k -mer size 4, using Random Forest algorithm. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 98.48%.

k -mer Value 5

For k -mer value 5, the results of the fine tuning indicated that the best performance was achieved with a number of estimators of 200, a maximum depth of 40 and bootstrapping. This configuration yielded an accuracy of **98.41%**, very close to the above accuracies. Detailed metrics, including the confusion matrix and classification report, are presented in Figure 3.12.

In the next step, the performance of the Random Forest algorithm in predicting Gentamicin resistance

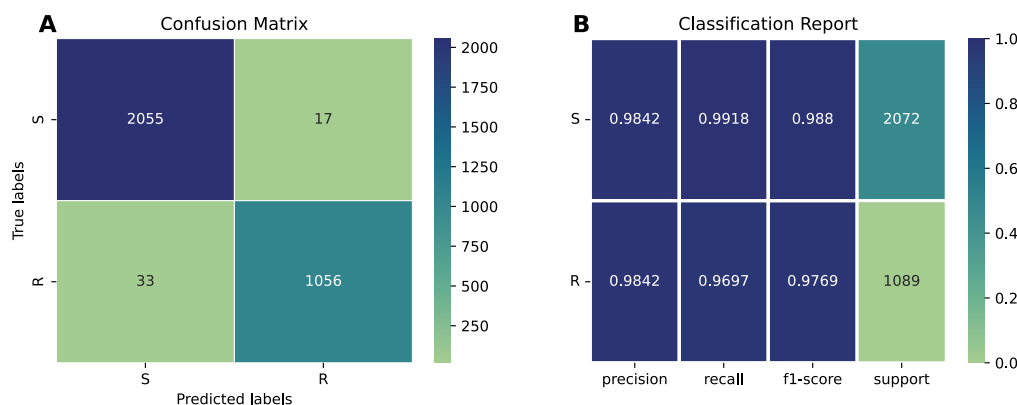


Figure 3.12 | Gentamicin predicted Phenotypes for k -mer size 5, using Random Forest algorithm. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 98.41%.

phenotypes was evaluated using two different encoding methods: one-hot encoding and PSSM encoding. The results for both encoding methods show high levels of accuracy, but with some differences in their detailed performance metrics.

For the Random Forest model with one-hot encoding (Figure 3.13), the overall accuracy achieved was 98.03%. The confusion matrix reveals that of the 2,080 susceptible samples, 2,053 were correctly classified, while 27 were misclassified as resistant. Similarly, of the 1,081 resistant samples, 1,046 were correctly classified, with 35 being misclassified as susceptible. The classification report indicates high precision, recall, and F1-scores for both classes: 0.9832 for precision, 0.987 for recall, and 0.9851 for F1-score in the susceptible class; 0.9748 for precision, 0.9676 for recall, and 0.9712 for F1-score in the resistant class. These metrics highlight the model's robustness in accurately distinguishing between susceptible and resistant strains using one-hot encoding. Comparing these results with the corresponding one of XGBoost (Figure 3.8), it turns out that the random forest has fewer false negatives and exactly the same accuracy, so we can argue that the random forest model makes a better prediction.

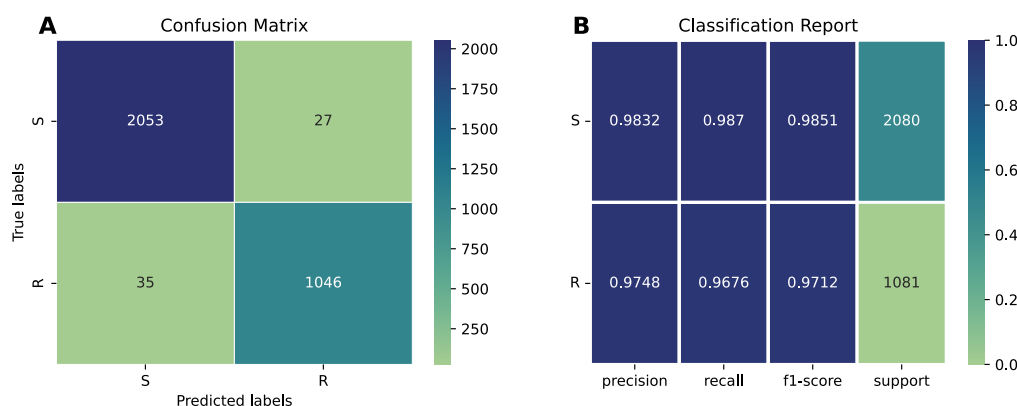


Figure 3.13 | Gentamicin predicted Phenotypes using Random Forest algorithm and one-hot encoding. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 98.03%.

In contrast, the Random Forest model that uses PSSM encoding (Figure 3.14) achieved a slightly lower overall accuracy of 97.78%. The confusion matrix shows that out of 2,072 susceptible samples, 2,047 were correctly identified and 25 were misclassified as resistant. For the 1,089 resistant samples, 1,044 were correctly classified, while 45 were misclassified as susceptible. The classification report for PSSM encoding indicates precision, recall, and F1-scores for the susceptible class as 0.9785, 0.988, and 0.983, respectively, and for the

resistant class as 0.976, 0.959, and 0.967. Although these metrics are lower than those achieved with one-hot encoding, they still demonstrate a strong predictive capability. Finally, comparing these results with those of XGBoost (Figure 3.9), it appears that the random forest has a slightly better accuracy and false negative rate.

In summary, both the one-hot and PSSM encoding methods provide high levels of accuracy in predicting Gentamicin resistance using the Random Forest algorithm. One-hot encoding slightly outperforms PSSM encoding in terms of overall accuracy and classification metrics. These results suggest that while PSSM encoding offers a robust k -mer alternative, one-hot encoding might be more effective for this particular application. However, the choice of encoding method can be influenced by other factors such as computational efficiency and the specific characteristics of the data set and finally the best accuracy was achieved with the k -mer data set and the random forest algorithm.

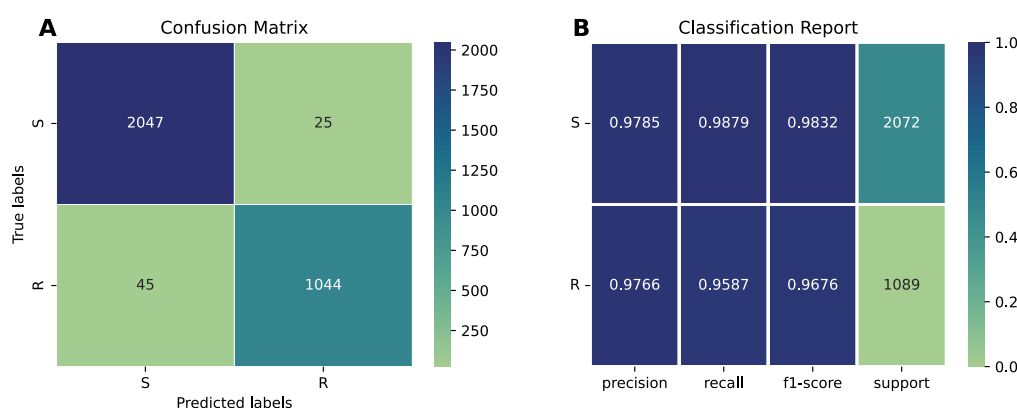


Figure 3.14 | Gentamicin predicted Phenotypes using Random Forest algorithm and PSSM encoding. (A) Confusion matrix of the model's predictions. (B) Classification report showcasing precision, recall, F1 score and support for each class. The predicted overall accuracy is 97.78%.

4 | DISCUSSION

In this study, we have employed machine learning to predict Antimicrobial Resistance (AMR) from Whole Genome Sequencing (WGS) data, focusing on seven common nosocomial pathogens, collectively known as the ESKAPEE pathogens: *Enterococcus*, *Staphylococcus*, *Klebsiella*, *Acinetobacter*, *Pseudomonas*, *Enterobacter species*, and *Escherichia*. These pathogens were selected for their clinical relevance and the abundance of AMR-annotated samples available in public databases, specifically the NCBI Pathogen Detection database (NDARO), the National Antimicrobial Resistance Monitoring System (NARMS), and the Bacterial and Viral Bioinformatics Resource Center (BV-BRC).

To build a comprehensive dataset, we collected data from the NCBI Pathogen Detection website, focusing on the ESKAPEE pathogens and filtering for Antimicrobial Susceptibility Testing (AST) phenotypes. This initial dataset included 13,827 isolates corresponding to 12,865 unique assemblies. Through the Entrez system, we retrieved antibiogram tables for these isolates, detailing the antibiotics tested, measurement values, units, and laboratory methods. We compiled these antibiograms into a single data frame, removing duplicates and records with inappropriate measurement units. Further, using a custom Snakemake pipeline, we downloaded missing assemblies and converted SRA DNA sequencing data, into assemblies using the SPAdes tool, resulting in 13,159 unique assemblies.

Similarly, data from the NARMS website was filtered and processed, yielding 477 unique BioSample IDs. We ensured consistency by transforming this data to align with other sources, removing irrelevant columns, and splitting the data by antibiotic. From the BV-BRC database, we downloaded AMR tables, filtered for relevant phenotypes and measurement units, and preprocessed the data to convert and download sequence files. Despite encountering technical challenges, such as unequal read numbers in some SRA files, we successfully processed the majority of records, achieving a final dataset of 6,249 unique assemblies.

By integrating data from NDARO, NARMS, and BV-BRC, we compiled a comprehensive data set of 19,885 records, representing 19,132 unique final assemblies after filtering duplicates and ensuring metadata completeness. This extensive and meticulously curated dataset formed the foundation for our machine learning models, ensuring a robust and representative sample for predicting AMR phenotypes. The applications of this expansive dataset are multiple and transformative, as it allows for the testing of various machine learning models on a large and well-characterized collection of data. This breadth and depth of data enable us to fine-tune and validate different algorithms, enhancing the accuracy and generalizability of AMR predictions. Furthermore, the diversity of the dataset ensures that our models are exposed to a wide range of genetic variations and resistance mechanisms, making them more resilient and reliable in real-world applications. As a result, this dataset is not only pivotal for advancing our current research but also holds immense potential for future studies on antimicrobial resistance and public health.

Finally, we applied quality control to the assemblies, based on the continuity value N50, which is also a quality indicator in them, and ended up with 18,545 assemblies, each of which we matched with an antibiogram.

The annotation pipeline was meticulously crafted to identify and catalog genes associated with antimicrobial resistance in the genomic assemblies. Initially, we assembled a comprehensive database of AMR genes using two primary resources: the Comprehensive Antibiotic Resistance Database (CARD) and the AMRFinderPlus. These databases provided extensive information on resistance genes, including sequences, variants, and mechanisms of resistance. Data from CARD was retrieved through the Antibiotic Resistance Ontology (ARO) table, while AMRFinderPlus data was accessed via the NCBI FTP site, resulting in a detailed gene catalog. We employed bash scripts to download and consolidate protein sequences from both databases, forming an Antimicrobial peptides (AMPs) database comprising core and plus proteins, which were subsequently clustered using CD-HIT to reduce redundancy and identify unique AMPs.

Following the creation of the AMPs database, we implemented a Snakemake pipeline to facilitate the

alignment of these sequences with WGS data. The CD-HIT tool was used to cluster the protein sequences, optimizing parameters to ensure accurate representation of AMR genes. The clustered results provided a basis for further analysis using the Basic Local Alignment Search Tool (BLAST), which identified the matching sites between the AMPs and assembly sequences. This process was refined to filter out duplicates and low-identity matches, ensuring high-quality data for downstream applications. Additionally, we incorporated ribosomal RNA (rRNA) information by aligning reference sequences of 5S, 16S, and 23S rRNAs with the genomic data, capturing non-translated genes targeted by certain antibiotics.

The inclusion of antimicrobial peptide sequences, their upstream sequences, and ribosomal RNA copies in our analysis serves several critical purposes in the prediction of Antimicrobial Resistance. Firstly, AMPs play a vital role in the innate immune response and are often the first line of defense against bacterial infections. By obtaining and clustering these sequences, we can identify specific AMPs and their variants that contribute to bacterial resistance mechanisms. These peptides interact with bacterial membranes or intracellular targets, and understanding their sequences helps us to pinpoint genetic markers associated with resistance. This information is essential for developing accurate models that can predict how different bacterial strains might respond to antimicrobial agents based on their genomic content.

Secondly, upstream sequences of these AMPs are crucial because they often contain regulatory elements that control gene expression. These regions can include promoters, operators, and other regulatory sequences that influence how AMPs are synthesized and function within the bacterial cell. By including upstream sequences in our analysis, we can gain insights into the regulatory mechanisms that can affect the expression levels of AMPs, thereby impacting the overall resistance phenotype of the bacterium. Understanding these regulatory elements allows us to predict not just the presence of resistance genes but also their potential activity within a bacterium.

Finally, incorporating rRNA sequences, particularly the 5S, 16S, and 23S rRNAs, is critical because many antibiotics target the bacterial ribosome, disrupting protein synthesis. rRNA genes are highly conserved and serve as reliable markers for phylogenetic studies and microbial identification. By aligning these rRNA sequences with our genomic data, we can identify mutations and variations that confer resistance to ribosome-targeting antibiotics. This comprehensive approach ensures that we capture a wide array of resistance mechanisms, from peptide interactions to ribosomal alterations, providing a holistic view of the genetic factors driving AMR. This multi-faceted dataset enhances our ability to identify resistance determinants and understand the complex interplay of genetic factors that contribute to the resistance of bacterial pathogens against antibiotics.

Central to our methodology was the application of machine learning models, specifically the XGBoost and the Random Forest algorithm, using different sequence encoding strategies, such as k -mer, one-hot and PSSM encoding. By performing hyperparameter tuning using various k -mer values and the aforementioned XGBoost objective functions, along with the hyperparameters detailed in Section 2.7.4, we achieved accuracy rates ranging from 97.62% up to 98.355%. Also, by performing hyperparameter tuning in the random forest model, again using various k -mer values, along with different hyperparameters, we achieved accuracy rates ranging from minimum 98.10% up to 98.48%. More metrics from these models are detailed below.

The Random Forest algorithm, when applied with k -mer encoding, exhibited an impressive overall accuracy of 98.48%. This high level of performance was evidenced by detailed metrics, including precision, recall, and F1-scores, which all underscored the model's ability to distinguish between susceptible and resistant strains effectively. Specifically, the model achieved a precision of 0.9842, recall of 0.992, and an F1-score of 0.9885 for susceptible strains, and a precision of 0.986, recall of 0.9697, and an F1-score of 0.9778 for resistant strains. Using the above algorithms with the other encodings, we achieved a slightly lower precision. More analytically, using one-hot coding, the models achieved exactly the same accuracy of 98.03%, while with PSSM encoding, the XGBoost model achieved accuracy of 97.63% and Random Forest 97.78%.

Our study achieved a maximum accuracy of 98.48% for predicting gentamicin resistance, surpassing the

results reported in several key articles in the literature. Van Camp *et al.*, 2020 reported an accuracy of 95.5%, while Moradigaravand *et al.*, 2018 achieved 97%. Nguyen *et al.*, 2018 and Nguyen *et al.*, 2019 documented accuracies of 95% and 91%, respectively, for gentamicin prediction. These previous studies typically focused on a limited scope, analyzing between 1 and 20 different antibiotics and covering from 1 to 5 different bacterial genera. In contrast, our research encompassed a far broader range, including 18,545 genomes from 7 bacterial genera, with a total dataset covering 97 different antibiotics and 14 combinations. This comprehensive approach not only yielded higher prediction accuracy, but also demonstrated the robustness and generalizability of our machine learning models across a diverse array of bacterial strains and antibiotics.

In conclusion, this research highlights the efficacy of machine learning in predicting AMR phenotypes from WGS data. These findings contribute significantly to the field of bioinformatics and AMR research, providing a foundation for future studies and the development of more precise diagnostic tools and targeted treatments. Continued exploration and refinement of these methodologies will be crucial in advancing our understanding and management of antimicrobial resistance.

Future implementations of our work will focus on expanding the scope of antimicrobial resistance prediction to include a wider range of antibiotics. By developing individual machine learning models tailored for each specific antibiotic, we aim to enhance the accuracy and reliability of resistance predictions across different antimicrobial agents. This approach will enable a more precise identification of resistance mechanisms for each antibiotic, facilitating targeted treatment strategies, and contributing to better clinical outcomes. Additionally, our methodology will be extended to cover a wider array of bacterial strains beyond the ESKAPEE pathogens, such as *Salmonella* and other clinically relevant bacteria. This generalization will make our predictive models more versatile and applicable to a broader spectrum of bacterial infections, enhancing their utility in diverse clinical settings.

Another promising direction involves incorporating Artificial Neural Networks (ANNs) into AMR prediction using our comprehensive dataset. ANNs, with their ability to model complex nonlinear relationships, offer the potential to uncover deeper insights into resistance mechanisms that might be missed by traditional machine learning algorithms. By leveraging the rich information contained in our annotated genomic data, ANNs can enhance the predictive power and robustness of our models. This integration of advanced neural network techniques will not only improve the accuracy of resistance predictions but also provide a more nuanced understanding of the genetic underpinnings of AMR. Overall, these future implementations will significantly advance our ability to predict and manage antimicrobial resistance, ultimately contributing to more effective public health strategies and better patient care.

5 | CONCLUSION

In conclusion, our study represents a significant advancement in the prediction of Antimicrobial Resistance (AMR) through the integration of a comprehensive and meticulously curated dataset of 18,545 unique final assemblies. This dataset, sourced from NDARO, NARMS, and BV-BRC databases, stands as the most extensive collection we have encountered for Machine learning (ML) training, offering a robust and representative sample for AMR phenotype prediction. The incorporation of Antimicrobial peptides (AMPs), their upstream sequences, and ribosomal RNA (rRNA) copies allowed us to capture a comprehensive genomic landscape, crucial for understanding the multifaceted mechanisms of resistance. Utilizing computational tools like Snakemake, CD-HIT, and BLAST, we ensured high-quality data processing, which facilitated the development of accurate and reliable ML models.

Our ML models, particularly those using XGBoost and random forest algorithms, achieved remarkable predictive accuracy, with a maximum accuracy of 98.48% for gentamicin resistance. This performance surpasses several notable studies and while previous research typically focused on a narrower scope, our work encompassed 18,545 genomes from 7 bacterial genera and covered 97 antibiotics and 14 combinations, demonstrating the robustness and scalability of our approach. The applications of this extensive dataset are manifold, enabling the testing of various ML models in a well-characterized and diverse dataset, which significantly enhances our capability to predict and manage AMR. Future directions include applying these predictive models to other antibiotics, incorporating artificial neural networks for even better performance, and generalizing the algorithm for additional bacterial strains beyond the ESKAPEE pathogens, ensuring broader applicability and utility in combating AMR.

6 | AVAILABILITY OF DATA

The data set and codes used in this thesis are publicly available and can be accessed through the following GitHub repository: [Machine Learning for Prediction of Antimicrobial Resistance](#). This repository includes links for this comprehensive data set compiled for the study, as well as the scripts and tools developed for the analysis, including the data acquisition pipeline, the annotation pipeline and the machine learning models. Researchers and practitioners are encouraged to utilize and build upon this work to further advance the field of antimicrobial resistance prediction.

BIBLIOGRAPHY

- [1] FLEMING, ALEXANDER (Jan. 1944). "THE DISCOVERY OF PENICILLIN". In: *British Medical Bulletin* 2.1, pp. 4–5. ISSN: 0007-1420. DOI: [10.1093/oxfordjournals.bmb.a071032](https://doi.org/10.1093/oxfordjournals.bmb.a071032). eprint: <https://academic.oup.com/bmb/article-pdf/2/1/4/6700621/2-1-4.pdf>. URL: <https://doi.org/10.1093/oxfordjournals.bmb.a071032>.
- [2] VENTOLA, C. LEE (2015). "The Antibiotic Resistance Crisis: Part 1: Causes and Threats". In: *Pharmacy and Therapeutics* 40.4, pp. 277–. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/>.
- [3] DAVIES, JULIAN and DOROTHY DAVIES (2010). "Origins and Evolution of Antibiotic Resistance". In: *Microbiology and Molecular Biology Reviews* 74.3, pp. 417–433. DOI: [10.1128/mmbr.00016-10](https://doi.org/10.1128/mmbr.00016-10). eprint: <https://journals.asm.org/doi/pdf/10.1128/mmbr.00016-10>. URL: <https://journals.asm.org/doi/abs/10.1128/mmbr.00016-10>.
- [4] WORLD HEALTH ORGANIZATION (2022). *Antimicrobial resistance surveillance in Europe 2022–2020 data*. World Health Organization. Regional Office for Europe. URL: <https://iris.who.int/bitstream/handle/10665/351141/9789289056687-eng.pdf>.
- [5] AMINOV, RUSTAM I. (2010). "A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future". In: *Frontiers in Microbiology* 1. ISSN: 1664-302X. DOI: [10.3389/fmicb.2010.00134](https://doi.org/10.3389/fmicb.2010.00134). URL: <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2010.00134>.
- [6] GOULD, KATE (Feb. 2016). "Antibiotics: from prehistory to the present day". In: *Journal of Antimicrobial Chemotherapy* 71.3, pp. 572–575. ISSN: 0305-7453. DOI: [10.1093/jac/dkv484](https://doi.org/10.1093/jac/dkv484). eprint: <https://academic.oup.com/jac/article-pdf/71/3/572/6994739/dkv484.pdf>. URL: <https://doi.org/10.1093/jac/dkv484>.
- [7] SCHATZ, ALBERT, ELIZABETH BUGIE, SELMAN A WAKSMAN, ARLEN D HANSSEN, ROBIN PATEL, and DOUGLAS R OSMON (Aug. 2005). "Streptomycin, a Substance Exhibiting Antibiotic Activity against Gram-Positive and Gram-Negative Bacteria". In: *Clinical Orthopaedics and Related Research* 437, pp. 3–6. ISSN: 0009-921X. DOI: [10.1097/01.blo.0000175887.98112.fe](https://doi.org/10.1097/01.blo.0000175887.98112.fe). URL: <http://dx.doi.org/10.1097/01.blo.0000175887.98112.fe>.
- [8] CASSINI, ALESSANDRO, LISELOTTE DIAZ HÖGBERG, DIAMANTIS PLACHOURAS, *et al.* (2019). "Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis". In: *The Lancet infectious diseases* 19.1, pp. 56–66.
- [9] WORLD HEALTH ORGANIZATION (2023). *Antimicrobial resistance surveillance in Europe 2023–2021 data*. World Health Organization. Regional Office for Europe. URL: <https://www.ecdc.europa.eu/en/publications-data/antimicrobial-resistance-surveillance-europe-2023-2021-data>.
- [10] STRATHDEE, STEFFANIE A, SALLY C DAVIES, and JASMINE R MARCELIN (Oct. 2020). "Confronting antimicrobial resistance beyond the COVID-19 pandemic and the 2020 US election". In: *The Lancet* 396.10257, pp. 1050–1053. ISSN: 0140-6736. DOI: [10.1016/S0140-6736\(20\)32063-8](https://doi.org/10.1016/S0140-6736(20)32063-8). URL: [http://dx.doi.org/10.1016/S0140-6736\(20\)32063-8](http://dx.doi.org/10.1016/S0140-6736(20)32063-8).
- [11] CENTERS FOR DISEASE CONTROL AND PREVENTION (2013). *Antibiotic Resistance Threats in the United States, 2013*. <https://www.cdc.gov/antimicrobial-resistance/media/pdfs/ar-threats-2013-508.pdf>. Accessed: 2024-05-25.
- [12] CENTERS FOR DISEASE CONTROL & PREVENTION (2019). *2019 Antibiotic Resistance Threats Report*. <https://www.cdc.gov/antimicrobial-resistance/media/pdfs/2019-ar-threats-report-508.pdf>. Accessed: 2024-05-22.
- [13] HSU, JEREMY (2020). "How covid-19 is accelerating the threat of antimicrobial resistance". In: *BMJ* 369. DOI: [10.1136/bmj.m1983](https://doi.org/10.1136/bmj.m1983). eprint: <https://www.bmj.com/content/369/bmj.m1983.full.pdf>. URL: <https://www.bmj.com/content/369/bmj.m1983>.
- [14] LANGFORD, BRADLEY J, MIRANDA SO, MARINA SIMEONOVA, *et al.* (Mar. 2023). "Antimicrobial resistance in patients with COVID-19: a systematic review and meta-analysis". In: *The Lancet Microbe* 4.3, e179–e191. ISSN: 2666-5247. DOI: [10.1016/S2666-5247\(22\)00355-X](https://doi.org/10.1016/S2666-5247(22)00355-X). URL: [http://dx.doi.org/10.1016/S2666-5247\(22\)00355-X](http://dx.doi.org/10.1016/S2666-5247(22)00355-X).
- [15] CENTERS FOR DISEASE CONTROL AND PREVENTION (2020). *COVID-19 & Antimicrobial Resistance | Antimicrobial Resistance | CDC*. <https://www.cdc.gov/antimicrobial-resistance/data-research/threats/covid-19.html>. Accessed: 2024-05-19.
- [16] NATIONAL CENTER FOR EMERGING AND ZOO NOTIC INFECTIOUS DISEASES (June 2022). *COVID-19: U.S. impact on antimicrobial resistance, special report 2022*. Division of Healthcare Quality Promotion. DOI: [10.15620/cdc.117915](https://doi.org/10.15620/cdc.117915). URL: <http://dx.doi.org/10.15620/cdc.117915>.

- [17] BRÖTZ-OESTERHELT, HEIKE and NINA A BRUNNER (2008). “How many modes of action should an antibiotic have?” In: *Current Opinion in Pharmacology* 8.5. Anti-infectives/New technologies, pp. 564–573. ISSN: 1471-4892. DOI: <https://doi.org/10.1016/j.coph.2008.06.008>. URL: <https://www.sciencedirect.com/science/article/pii/S1471489208000799>.
- [18] HODGKIN, DC (1949). “The X-ray analysis of the structure of penicillin”. In: *Advancement of science* 6.22. URL: <https://pubmed.ncbi.nlm.nih.gov/18134678/>.
- [19] SHEEHAN, JOHN C. and KENNETH R. HENERY-LOGAN (1959). “The Total Synthesis of Penicillin V”. In: *Journal of the American Chemical Society* 81.12, pp. 3089–3094. DOI: [10.1021/ja01521a044](https://doi.org/10.1021/ja01521a044). eprint: <https://doi.org/10.1021/ja01521a044>. URL: <https://doi.org/10.1021/ja01521a044>.
- [20] SPRATT, B G (1975). “Distinct penicillin binding proteins involved in the division, elongation, and shape of Escherichia coli K12.” In: *Proceedings of the National Academy of Sciences* 72.8, pp. 2999–3003. DOI: [10.1073/pnas.72.8.2999](https://doi.org/10.1073/pnas.72.8.2999). eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.72.8.2999>. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.72.8.2999>.
- [21] TIPPER, D J and J L STROMINGER (1965). “Mechanism of action of penicillins: a proposal based on their structural similarity to acyl-D-alanyl-D-alanine.” In: *Proceedings of the National Academy of Sciences* 54.4, pp. 1133–1141. DOI: [10.1073/pnas.54.4.1133](https://doi.org/10.1073/pnas.54.4.1133). eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.54.4.1133>. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.54.4.1133>.
- [22] DAVIES, J and B D DAVIS (1968). “Misreading of Ribonucleic Acid Code Words Induced by Aminoglycoside Antibiotics: THE EFFECT OF DRUG CONCENTRATION”. In: *Journal of Biological Chemistry* 243.12, pp. 3312–3316. ISSN: 0021-9258. DOI: [https://doi.org/10.1016/S0021-9258\(18\)93308-9](https://doi.org/10.1016/S0021-9258(18)93308-9). URL: <https://www.sciencedirect.com/science/article/pii/S0021925818933089>.
- [23] DRLICA, K and X ZHAO (1997). “DNA gyrase, topoisomerase IV, and the 4-quinolones”. In: *Microbiology and Molecular Biology Reviews* 61.3, pp. 377–392. DOI: [10.1128/mmbr.61.3.377-392.1997](https://doi.org/10.1128/mmbr.61.3.377-392.1997). URL: <https://journals.asm.org/doi/abs/10.1128/mmbr.61.3.377-392.1997>.
- [24] WEHRLI, W and M STAEHELIN (1971). “Actions of the rifamycins”. In: *Bacteriological Reviews* 35.3, pp. 290–309. DOI: [10.1128/br.35.3.290-309.1971](https://doi.org/10.1128/br.35.3.290-309.1971). URL: <https://journals.asm.org/doi/abs/10.1128/br.35.3.290-309.1971>.
- [25] BROWN, GENE M. (1962). “The Biosynthesis of Folic Acid: II. INHIBITION BY SULFONAMIDES”. In: *Journal of Biological Chemistry* 237.2, pp. 536–540. ISSN: 0021-9258. DOI: [https://doi.org/10.1016/S0021-9258\(18\)93957-8](https://doi.org/10.1016/S0021-9258(18)93957-8). URL: <https://www.sciencedirect.com/science/article/pii/S0021925818939578>.
- [26] EDWARDS, DAVID I. (Jan. 1993). “Nitroimidazole drugs-action and resistance mechanisms I. Mechanism of action”. In: *Journal of Antimicrobial Chemotherapy* 31.1, pp. 9–20. ISSN: 0305-7453. DOI: [10.1093/jac/31.1.9](https://doi.org/10.1093/jac/31.1.9). eprint: <https://academic.oup.com/jac/article-pdf/31/1/9/2398443/31-1-9.pdf>. URL: <https://doi.org/10.1093/jac/31.1.9>.
- [27] DJORDJEVIC, STEVEN P., VERONICA M. JAROCKI, TORSTEN SEEMANN, *et al.* (Sept. 2023). “Genomic surveillance for antimicrobial resistance — a One Health perspective”. In: *Nature Reviews Genetics* 25.2, pp. 142–157. ISSN: 1471-0064. DOI: [10.1038/s41576-023-00649-y](https://doi.org/10.1038/s41576-023-00649-y). URL: <http://dx.doi.org/10.1038/s41576-023-00649-y>.
- [28] WOLDEGEORGIS, BESHADA ZERFU, AMENE ABEBE KERBO, MOHAMMED SULEIMAN OBSA, and TAKLU MARAMA MOKONNON (June 2023). “A systematic review and meta-analysis of antimicrobial resistance knowledge, attitudes, and practices: Current evidence to build a strong national antimicrobial drug resistance narrative in Ethiopia”. In: *PLOS ONE* 18.6, pp. 1–18. DOI: [10.1371/journal.pone.0287042](https://doi.org/10.1371/journal.pone.0287042). URL: <https://doi.org/10.1371/journal.pone.0287042>.
- [29] MARTINEZ, JOSE L. (2009). “The role of natural environments in the evolution of resistance traits in pathogenic bacteria”. In: *Proceedings of the Royal Society B: Biological Sciences* 276.1667, pp. 2521–2530. DOI: [10.1098/rspb.2009.0320](https://doi.org/10.1098/rspb.2009.0320). URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2009.0320>.
- [30] BOUCHER, HELEN W., GEORGE H. TALBOT, JOHN S. BRADLEY, *et al.* (Jan. 2009). “Bad Bugs, No Drugs: No ESKAPE! An Update from the Infectious Diseases Society of America”. In: *Clinical Infectious Diseases* 48.1, pp. 1–12. ISSN: 1058-4838. DOI: [10.1086/595011](https://doi.org/10.1086/595011). eprint: <https://academic.oup.com/cid/article-pdf/48/1/1/812476/48-1-1.pdf>. URL: <https://doi.org/10.1086/595011>.

- [31] SANTAJIT, SIRIJAN and NITAYA INDRAWATTANA (2016). “Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens”. In: *BioMed Research International* 2016.1, p. 2475067. DOI: <https://doi.org/10.1155/2016/2475067>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2016/2475067>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2016/2475067>.
- [32] MUNITA, JOSE M. and CESAR A. ARIAS (2016). “Mechanisms of Antibiotic Resistance”. In: *Microbiology Spectrum* 4.2, 10.1128/microbiolspec.vmbf-0016–2015. DOI: [10.1128/microbiolspec.vmbf-0016–2015](https://doi.org/10.1128/microbiolspec.vmbf-0016-2015). URL: <https://journals.asm.org/doi/abs/10.1128/microbiolspec.vmbf-0016-2015>.
- [33] TACCONELLI, EVELINA, ELENA CARRARA, ALESSIA SAVOLDI, *et al.* (2018). “Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis”. In: *The Lancet Infectious Diseases* 18.3, pp. 318–327. ISSN: 1473-3099. DOI: [https://doi.org/10.1016/S1473-3099\(17\)30753-3](https://doi.org/10.1016/S1473-3099(17)30753-3). URL: <https://www.sciencedirect.com/science/article/pii/S1473309917307533>.
- [34] WORLD HEALTH ORGANIZATION (2014). *Antimicrobial resistance: global report on surveillance*. <https://www.who.int/publications/i/item/9789241564748>. Accessed: 2024-06-12.
- [35] POOLE, K (Apr. 2001). “Multidrug efflux pumps and antimicrobial resistance in *Pseudomonas aeruginosa* and related organisms”. In: *Journal of molecular microbiology and biotechnology* 3.2, pp. 255–264. ISSN: 1464-1801. URL: <http://europepmc.org/abstract/MED/11321581>.
- [36] CHAMBERS, HENRY F. and FRANK R. DELEO (Sept. 2009). “Waves of resistance: *Staphylococcus aureus* in the antibiotic era”. In: *Nature Reviews Microbiology* 7.9, pp. 629–641. ISSN: 1740-1534. DOI: [10.1038/nrmicro2200](https://doi.org/10.1038/nrmicro2200). URL: <http://dx.doi.org/10.1038/nrmicro2200>.
- [37] PARTRIDGE, SALLY R., STEPHEN M. KWONG, NEVILLE FIRTH, and SLADE O. JENSEN (2018). “Mobile Genetic Elements Associated with Antimicrobial Resistance”. In: *Clinical Microbiology Reviews* 31.4, 10.1128/cmr.00088–17. DOI: [10.1128/cmr.00088–17](https://doi.org/10.1128/cmr.00088-17). URL: <https://journals.asm.org/doi/abs/10.1128/cmr.00088-17>.
- [38] RELLER, L. BARTH, MELVIN WEINSTEIN, JAMES H. JORGENSEN, and MARY JANE FERRARO (Dec. 2009). “Antimicrobial Susceptibility Testing: A Review of General Principles and Contemporary Practices”. In: *Clinical Infectious Diseases* 49.11, pp. 1749–1755. ISSN: 1058-4838. DOI: [10.1086/647952](https://doi.org/10.1086/647952). eprint: <https://academic.oup.com/cid/article-pdf/49/11/1749/984933/49-11-1749.pdf>. URL: <https://doi.org/10.1086/647952>.
- [39] TURNIDGE, JOHN and DAVID L. PATERSON (2007). “Setting and Revising Antibacterial Susceptibility Breakpoints”. In: *Clinical Microbiology Reviews* 20.3, pp. 391–408. DOI: [10.1128/cmr.00047–06](https://doi.org/10.1128/cmr.00047-06). URL: <https://journals.asm.org/doi/abs/10.1128/cmr.00047-06>.
- [40] ANDREWS, JENNIFER M. (July 2001). “Determination of minimum inhibitory concentrations”. In: *Journal of Antimicrobial Chemotherapy* 48.suppl_1, pp. 5–16. ISSN: 0305-7453. DOI: [10.1093/jac/48.suppl_1.5](https://doi.org/10.1093/jac/48.suppl_1.5). URL: https://doi.org/10.1093/jac/48.suppl_1.5.
- [41] WIEGAND, IRITH, KAI HILPERT, and ROBERT E W HANCOCK (Jan. 2008). “Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances”. In: *Nature Protocols* 3.2, pp. 163–175. ISSN: 1750-2799. DOI: [10.1038/nprot.2007.521](https://doi.org/10.1038/nprot.2007.521). URL: <http://dx.doi.org/10.1038/nprot.2007.521>.
- [42] CLSI (2020). *Analysis and Presentation of Cumulative Antimicrobial Susceptibility Test Data, 4th Edition (M39)*. https://clsi.org/media/1454/m39a4_sample.pdf. Clinical and Laboratory Standards Institute. Accessed: 2024-06-13.
- [43] NCBI (2023). *National Database of Antibiotic Resistant Organisms (NDARO)*. Accessed: 2024-05-24. URL: <https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/>.
- [44] OLSON, ROBERT D, RIDA ASSAF, THOMAS BRETTIN, *et al.* (Nov. 2022). “Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR”. In: *Nucleic Acids Research* 51.D1, pp. D678–D689. ISSN: 0305-1048. DOI: [10.1093/nar/gkac1003](https://doi.org/10.1093/nar/gkac1003). eprint: <https://academic.oup.com/nar/article-pdf/51/D1/D678/48441350/gkac1003.pdf>. URL: <https://doi.org/10.1093/nar/gkac1003>.
- [45] KARP, BETH E., HEATHER TATE, JODIE R. PLUMBLEE, *et al.* (2017). “National Antimicrobial Resistance Monitoring System: Two Decades of Advancing Public Health Through Integrated Surveillance of Antimicrobial Resistance”. In: *Foodborne Pathogens and Disease* 14.10. PMID: 28792800, pp. 545–557. DOI: [10.1089/fpd.2017.2283](https://doi.org/10.1089/fpd.2017.2283). URL: <https://doi.org/10.1089/fpd.2017.2283>.

- [46] NGUYEN, MARCUS, THOMAS BRETTIN, S. WESLEY LONG, *et al.* (Jan. 2018). “Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*”. In: *Scientific Reports* 8.1. ISSN: 2045-2322. DOI: [10.1038/s41598-017-18972-w](https://doi.org/10.1038/s41598-017-18972-w). URL: <http://dx.doi.org/10.1038/s41598-017-18972-w>.
- [47] KIM, JEE IN, FINLAY MAGUIRE, KARA K. TSANG, THEODORE GOULIOURIS, SHARON J. PEACOCK, TIM A. MCALLISTER, ANDREW G. MCARTHUR, and ROBERT G. BEIKO (2022). “Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective”. In: *Clinical Microbiology Reviews* 35.3, e00179–21. DOI: [10.1128/cmr.00179-21](https://doi.org/10.1128/cmr.00179-21). URL: <https://journals.asm.org/doi/abs/10.1128/cmr.00179-21>.
- [48] MARTÍNEZ-AGÜERO, SERGIO, INMACULADA MORA-JIMÉNEZ, JON LÉRIDA-GARCÍA, JOAQUÍN ÁLVAREZ-RODRÍGUEZ, and CRISTINA SOGUERO-RUIZ (2019). “Machine Learning Techniques to Identify Antimicrobial Resistance in the Intensive Care Unit”. In: *Entropy* 21.6. ISSN: 1099-4300. DOI: [10.3390/e21060603](https://doi.org/10.3390/e21060603). URL: <https://www.mdpi.com/1099-4300/21/6/603>.
- [49] MACESIC, NENAD, FERNANDA POLUBRIAGINOF, and NICHOLAS P. TATONETTI (Dec. 2017). “Machine learning: novel bioinformatics approaches for combating antimicrobial resistance”. In: *Current Opinion in Infectious Diseases* 30.6, pp. 511–517. ISSN: 1473-6527. DOI: [10.1097/qco.0000000000000406](https://doi.org/10.1097/qco.0000000000000406). URL: <http://dx.doi.org/10.1097/QCO.0000000000000406>.
- [50] KHALEDI, ARIANE, AARON WEIMANN, MONIKA SCHNIEDERJANS, *et al.* (2020). “Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics”. In: *EMBO Molecular Medicine* 12.3, e10264. DOI: <https://doi.org/10.15252/emmm.201910264>. URL: <https://www.embopress.org/doi/abs/10.15252/emmm.201910264>.
- [51] MORADIGARAVAND, DANESH, MARTIN PALM, ANNE FAREWELL, VILLE MUSTONEN, JONAS WARRINGER, and LEOPOLD PARTS (Dec. 2018). “Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data”. In: *PLOS Computational Biology* 14.12, pp. 1–17. DOI: [10.1371/journal.pcbi.1006258](https://doi.org/10.1371/journal.pcbi.1006258). URL: <https://doi.org/10.1371/journal.pcbi.1006258>.
- [52] REN, YUNXIAO, TRINAD CHAKRABORTY, SWAPNIL DOJAD, *et al.* (Oct. 2021). “Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning”. In: *Bioinformatics* 38.2, pp. 325–334. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btab681](https://doi.org/10.1093/bioinformatics/btab681). URL: <https://doi.org/10.1093/bioinformatics/btab681>.
- [53] VAN CAMP, PIETER-JAN, DAVID B. HASLAM, and ALEKSEY POROLLO (2020). “Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data”. In: *Frontiers in Microbiology* 11. ISSN: 1664-302X. DOI: [10.3389/fmicb.2020.01013](https://doi.org/10.3389/fmicb.2020.01013). URL: <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2020.01013>.
- [54] NGUYEN, MARCUS, S. WESLEY LONG, PATRICK F. MCDERMOTT, *et al.* (2019). “Using Machine Learning To Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal *Salmonella*”. In: *Journal of Clinical Microbiology* 57.2, 10.1128/jcm.01260–18. DOI: [10.1128/jcm.01260-18](https://doi.org/10.1128/jcm.01260-18). URL: <https://journals.asm.org/doi/abs/10.1128/jcm.01260-18>.
- [55] SCHULER, GREGORY D., JONATHAN A. EPSTEIN, HITOMI OHKAWA, and JONATHAN A. KANS (1996). “Entrez: Molecular biology database and retrieval system”. In: *Computer Methods for Macromolecular Sequence Analysis*. Vol. 266. Methods in Enzymology. Academic Press, pp. 141–162. DOI: [https://doi.org/10.1016/S0076-6879\(96\)66012-1](https://doi.org/10.1016/S0076-6879(96)66012-1). URL: <https://www.sciencedirect.com/science/article/pii/S0076687996660121>.
- [56] MÖLDER, FELIX, KIM PHILIPP JABLONSKI, BRICE LETCHER, *et al.* (Apr. 2021). “Sustainable data analysis with Snakemake”. In: *F1000Research* 10, p. 33. ISSN: 2046-1402. DOI: [10.12688/f1000research.29032.2](https://doi.org/10.12688/f1000research.29032.2). URL: <http://dx.doi.org/10.12688/f1000research.29032.2>.
- [57] BANKEVICH, ANTON, SERGEY NURK, DMITRY ANTIPOV, *et al.* (2012). “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing”. In: *Journal of Computational Biology* 19.5. PMID: 22506599, pp. 455–477. DOI: [10.1089/cmb.2012.0021](https://doi.org/10.1089/cmb.2012.0021). eprint: <https://doi.org/10.1089/cmb.2012.0021>. URL: <https://doi.org/10.1089/cmb.2012.0021>.
- [58] PAGÈS, HERVÉ, PATRICK ABOYOUN, ROBERT GENTLEMAN, and SAIKAT DEBROY (2024). *Biostrings: Efficient manipulation of biological strings*. R package version 2.72.1. URL: <https://bioconductor.org/packages/Biostrings>.

- [59] NARAYANAN, NAVANEETH and JAMES S LEWIS (2023). “Aminoglycoside Therapeutic Drug Monitoring: On Paper vs in Practice”. In: *Clinical Infectious Diseases* 77.12, pp. 1737–1738.
- [60] ALCOCK, BRIAN P, WILLIAM HUYNH, ROMEO CHALIL, *et al.* (Oct. 2022). “CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database”. In: *Nucleic Acids Research* 51.D1, pp. D690–D699. ISSN: 0305-1048. DOI: [10.1093/nar/gkac920](https://doi.org/10.1093/nar/gkac920). eprint: <https://academic.oup.com/nar/article-pdf/51/D1/D690/48441073/gkac920.pdf>. URL: <https://doi.org/10.1093/nar/gkac920>.
- [61] FELDGARDEN, MICHAEL, VYACHESLAV BROVER, NARJOL GONZALEZ-ESCALONA, *et al.* (June 2021). “AMRFinder-Plus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence”. In: *Scientific Reports* 11.1. ISSN: 2045-2322. DOI: [10.1038/s41598-021-91456-0](https://doi.org/10.1038/s41598-021-91456-0). URL: <http://dx.doi.org/10.1038/s41598-021-91456-0>.
- [62] FU, LIMIN, BEIFANG NIU, ZHENGWEI ZHU, SITAO WU, and WEIZHONG LI (Oct. 2012). “CD-HIT: accelerated for clustering the next-generation sequencing data”. In: *Bioinformatics* 28.23, pp. 3150–3152. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bts565](https://doi.org/10.1093/bioinformatics/bts565). eprint: https://academic.oup.com/bioinformatics/article-pdf/28/23/3150/50695216/bioinformatics_28_23_3150.pdf. URL: <https://doi.org/10.1093/bioinformatics/bts565>.
- [63] ZILHAO, R, B PAPADOPOULOU, and P COURVALIN (1988). “Occurrence of the Campylobacter resistance gene tetO in Enterococcus and Streptococcus spp”. In: *Antimicrobial Agents and Chemotherapy* 32.12, pp. 1793–1796. DOI: [10.1128/aac.32.12.1793](https://doi.org/10.1128/aac.32.12.1793). eprint: <https://journals.asm.org/doi/pdf/10.1128/aac.32.12.1793>. URL: <https://journals.asm.org/doi/abs/10.1128/aac.32.12.1793>.
- [64] DEVIKA, S, L JEYASEELAN, and G SEBASTIAN (2016). “Analysis of sparse data in logistic regression in medical research: A newer approach”. In: *Journal of Postgraduate Medicine* 62.1, pp. 26–31. ISSN: 0022-3859. DOI: [10.4103/0022-3859.173193](https://doi.org/10.4103/0022-3859.173193). URL: <http://dx.doi.org/10.4103/0022-3859.173193>.
- [65] CLAUSEN, PHILIP T. L. C., FRANK M. AARESTRUP, and OLE LUND (Aug. 2018). “Rapid and precise alignment of raw reads against redundant databases with KMA”. In: *BMC Bioinformatics* 19.1. ISSN: 1471-2105. DOI: [10.1186/s12859-018-2336-6](https://doi.org/10.1186/s12859-018-2336-6). URL: <http://dx.doi.org/10.1186/s12859-018-2336-6>.
- [66] PEDREGOSA, F., G. VAROQUAUX, A. GRAMFORT, *et al.* (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- [67] HARRIS, CHARLES R., K. JARROD MILLMAN, STÉFAN J. VAN DER WALT, *et al.* (Sept. 2020). “Array programming with NumPy”. In: *Nature* 585.7825, pp. 357–362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). URL: <https://doi.org/10.1038/s41586-020-2649-2>.
- [68] SIEVERS, FABIAN and DESMOND G. HIGGINS (2018). “Clustal Omega for making accurate alignments of many protein sequences”. In: *Protein Science* 27.1, pp. 135–145. DOI: <https://doi.org/10.1002/pro.3290>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3290>.
- [69] CHEN, TIANQI and CARLOS GUESTRIN (2016). “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, pp. 785–794. ISBN: 9781450342322. DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785). URL: <https://doi.org/10.1145/2939672.2939785>.
- [70] BREIMAN, LEO (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 0885-6125. DOI: [10.1023/a:1010933404324](https://doi.org/10.1023/a:1010933404324).
- [71] KOHAVI, RON (1995). “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI’95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 1137–1143. ISBN: 1558603638.
- [72] JAPKOWICZ, NATHALIE and SHAJU STEPHEN (Oct. 2002). “The class imbalance problem: A systematic study”. In: *Intell. Data Anal.* 6.5, pp. 429–449. ISSN: 1088-467X.